

A Multi-tiered Approach for Analyzing Expressive Timing in Music Performance

Panayotis Mavromatis

Department of Music and Performing Arts,
New York University,
35 West 4th St., Suite 777,
New York, NY 10012, USA
panos.mavromatis@nyu.edu
<http://theory.smusic.nyu.edu/pm/>

Abstract. This paper presents a method for analyzing expressive timing data from music performances. The goal is to uncover rules which explain a performer's systematic timing manipulations in terms of structural features of the music such as form, harmonic progression, texture, and rhythm. A multi-tiered approach is adopted, in which one first identifies a continuous *tempo curve* by performing non-linear regression on the durations of performed time spans at all levels in the metric hierarchy. Once the effect of tempo has been factored out, subsequent tiers of analysis examine how the performed subdivision of each metric layer (e.g., quarter note) typically deviates from an even rendering of the next lowest layer (e.g., two equal eighth notes) as a function of time. Structural features in the music are identified that contribute to a performer's tempo fluctuations and metric deviations.

1 Introduction

The study of expressive musical performance has been the subject of experimental as well as computational research [1,2]. It is generally acknowledged that expressive timing—a performer's deviations from an exact temporal rendering of the score—is an important component of musical expression. By manipulating timing, a performer is able to communicate musical structure and shape a listener's experience of the music. This paper presents a method for analyzing expressive timing data, extracted through audio analysis of recorded performances. The purpose is to uncover rules which explain a performer's systematic timing manipulations in terms of structural features such as form, harmonic progression, texture, and rhythm.

A fundamental assumption of this analysis is that a performer controls a hierarchically structured *metrical cycle* of measure, beat, and subdivision levels [3,4]. At each point in time, the performer's mental clock fires at a given tempo, which is evidenced by the *cumulative* effect of all the levels in the metrical cycle. The performer's clock rate as a function of time is represented by a *tempo curve*. Identifying this curve forms a natural first tier of analysis.

Once the effect of tempo has been factored out, it is possible to examine how the performance rendering of subdivisions between adjacent metrical layers (e.g., the subdivision of a quarter note into two eighth notes) deviates from the corresponding exact duration ratios (e.g., 0.5 / 0.5). In the subsequent tiers of analysis, systematic deviations of this type are identified at each level in the metric hierarchy.

As justification for the proposed multi-tiered approach, we first note that it is supported by informal musical discourse: terms such as *ritardando* and *accelerando* typically refer to the first tier of expressive timing, whereas terms such as *rubato*, *notes inégales*, or “swing” most commonly represent deviations in the subsequent tiers.

More to the point, it appears that, in principle at least, a skilled performer can control each tier independently. For instance, a performer may be asked to manipulate the tempo of a performance, while maintaining even metric subdivisions. Conversely, the performer may be requested to perform at steady tempo, while producing various types of uneven metric subdivisions. Moreover, these uneven subdivisions can be executed independently at any particular metric level—up to a certain depth—while maintaining even timing at higher metrical levels.

One of the challenges of expressive performance research is to understand the cognitive mechanisms that underlie expert expressive rendering of a musical score. In line with current views on cognitive modeling, it is natural to seek *modular* rules that specialize in responding to specific features of the musical structure (such as metric accent) by shaping the expression in specific ways (such as lengthening the strong half of a subdivision). This modularity requirement poses challenges for any analytical approach to expressive performance data, which must identify and isolate the effect of individual rules from their surrounding context, where other rules may be simultaneously contributing to expressive deviations. It is in this spirit that the present analysis is offered; it represents work aiming towards a complete rule system for expressive timing. We believe that the multi-tiered analytical approach proposed in this paper can help identify and isolate the right ingredients in this complex multi-faceted manifestation of expert musical skill.

2 Related Previous Research

Several studies have focused on modeling specific aspects of expressive performance, such as *rubato* [5] or the final *ritardando* (see [6] for a review). In addition, some research groups have aimed for comprehensive models that integrate many different components of performance expression. As expected, timing plays a central role in such models.

An important early attempt at an integrated model was the work of Eric Clarke [7]. Clarke proposed nine generative rules to explain expressive deviations in terms of the performed piece’s structural features, such as grouping and meter. These rules were derived from measurements of piano performances in

experimental studies by Clarke and collaborators. Another important contribution was the KTH model by Sundberg and his group [8]. This model represents a synthetic approach, where expression rules were formulated by querying expert performers as to their expressive deviation practices.

The approach proposed in the present paper is inspired in part by the work of Gerhard Widmer and his collaborators [9,10,2], perhaps the most sophisticated proposal to date for an integrated model of expressive performance. Widmer's group applied machine learning techniques to analyze measurements of expressive performance by skilled musicians.

Most relevant to our approach is the fact that Widmer employed a *two-tiered* data analytic model, in which local note-to-note expressive deviations were separated from the more global expressive shaping of grouping units, such as phrases. Following earlier research [11], Widmer hypothesized that each grouping unit in the music contributes a parabolically shaped *accelerando-ritardando* component to the performance's tempo curve. The overall tempo curve is assumed to be the product of all such contributions coming from each grouping unit.

The first tier of Widmer's analysis consisted of identifying the parabolic coefficients corresponding to each unit of grouping. The process started from the highest grouping level and proceeded to the lowest. At each level, the coefficients were identified by least-squares fitting. After each level's contribution was factored out, the analysis was repeated at the next lowest level, until all levels of grouping were accounted for. The residual timing deviations were then attributed to local note-to-note expressive timing rules, which were extracted from the data via a machine learning algorithm.

The present work extends and modifies Widmer's approach in two different ways. First, we do not make the assumption that grouping is the only factor contributing to the shape of the tempo curve. Instead, we consider sources of additional contributions, such as texture and the tonal/formal function of phrases and sections. As we will see, there is indeed evidence that such factors come into play in determining a performance's tempo fluctuations.

The second difference between Widmer's approach and ours is that, rather than examining a single layer of low-level residual timing deviations, we analyze *separately* the deviations at each subdivision level in the metric hierarchy. As we will see, there is evidence that this separation could lead to simpler, more modular rules. At the same time, this allows us to develop rules that are specific to the absolute time scale of metric subdivision, as measured in seconds. Indeed, different time scales of pulsation can have different cognitive properties, as evidenced by several experimental studies, which are nicely summarized in [4] (see especially Chapter 2).

3 Tempo Curve Calculation Using a Non-Parametric Regression Model

If we give up a specific functional dependence of the tempo curve on grouping, as implemented by the fitting of parabolic segments, we must consider the most

general options for calculating the tempo curve from the timing data. This naturally leads to a non-parametric regression analysis, which does not assume a specific functional form for the tempo curve.

For the purposes of this study, we found it most flexible to use a non-linear regression model based on *radial basis functions*. The technique was first proposed in [12,13], and is a particular instance of density estimation using *Parzen windows* [14]. In its simplest form, the process can be illustrated as follows:

Let $\{x_i : i = 1 \dots N\}$ be a set of values for the independent variable X , and let $\{y_i : i = 1 \dots N\}$ be the corresponding values of the dependent variable Y , so that (x_i, y_i) are the coordinates of the i -th point in the data set. Then the regression curve $y(x)$ obtained from the above data set is given by

$$y(x) = \frac{\sum_{i=1}^N y_i \exp[-(x - x_i)^2/2\sigma^2]}{\sum_{i=1}^N \exp[-(x - x_i)^2/2\sigma^2]}$$

under the assumption of a Gaussian Parzen window. This expression, calculated through the Parzen density estimation formula, has a simple interpretation: it tells us that the predicted value of y at point x is equal to a weighted sum of the y_i observed at each x_i . The weights are determined by the distance of each x_i from x , and decay rapidly with that distance, according to a Gaussian function.

The variance σ of the Parzen window is also known as the *window width*, and can be viewed as a kind of smoothing parameter. Thus, the regression is formally equivalent to a (Gaussian-weighted) moving average filter. However, it should be noted that the window width is not set a priori, but it is inferred from the data. Indeed, a central problem of this regression analysis is to determine the right value of σ : If the latter is too large, the regression curve becomes too coarse to capture the meaningful fluctuations in the data. Conversely, when σ is too small, the regression curve displays over-fitting, i.e., it captures random noise fluctuations in the data and is a poor predictive model. A simple, yet effective way to determine the appropriate value of σ is through a form of *N-fold cross-validation*. This is effected by minimizing a cost function that represents a least-squares error on the cross-validation training sets (see [13] for more details).

The starting point for our analysis is a performance's set of *inter-onset durations*. These are extracted from the audio recording using Tristan Jehan's Echo Nest API. The latter is a programming toolkit for digital audio analysis that contains a tool for automatic note onset detection¹. Depending on the value of a resolution parameter, the algorithm can miss a real note onset (if the resolution is too low), or detect a spurious one (e.g. caused by reverb, if the resolution is too high). There is no single optimal resolution, and so it is generally safest to perform onset detection using a relatively high value, to ensure that no notes have been missed. As a result, any spurious onsets detected by the algorithm must be filtered out manually by listening. The algorithm produces the time of each onset in seconds, correct to four decimal places, from which the inter-onset duration values can be calculated at the same precision.

¹ See <http://developer.echonest.com/pages/overview> (last visited March 2009)

Since each note's inter-onset duration reflects not only the local tempo, but also the note's *nominal* duration value (e.g., quarter-note, eighth note, etc.), we must normalize each of the raw inter-onset durations by dividing it by the corresponding note's nominal value, where a whole note equals 1.0, a quarter note equals 0.25, etc. This way, each normalized inter-onset duration is a consistent indicator of the local tempo: its value reflects the whole-note duration corresponding to the tempo at that specific point in time. Our solution is essentially equivalent to Widmer's representation of his timing data using percentage deviations instead of absolute durations, but has the added advantage that it keeps track of *absolute* tempo information, and not just its relation to some average. The normalized inter-onset durations for each performance were used as data presented to the non-linear regression model, in order to obtain that performance's tempo curve.

Figure 1 shows the application of the above analysis to a recording of Bach's F minor prelude, BWV 881, from the Well-Tempered Clavier, Book 2. The piece is performed on the harpsichord by an expert, and is recorded on a commercially available CD. This performance will be used as an illustration throughout the paper. In Figure 1, the data points corresponding to the normalized inter-onset durations are shown in grey. The tempo curve derived from the regression is shown in black.

The performances of three contrasting Bach preludes (BWV 845, 863, 881) were analyzed, each of them performed by two different harpsichordists. The most salient factors shaping the tempo curve appear to be

- an initial small *accelerando*;
- a pronounced final *ritardando*;
- less pronounced, but consistent *ritardandi* leading to important cadences, with magnitude usually reflecting the cadence's hierarchical depth in the Schenkerian sense;
- small but measurable contrasts in tempo to highlight sections marked off by distinctive texture or tonal function (e.g., extended dominant pedal).

Once the effect of tempo is factored out, one can examine the lengthening or shortening of individual measures with respect to their neighbors, in response to specific features of the music. This individual manipulation of measure lengths is distinct from overall tempo change, and can be represented in a graph such as that of Fig. 2. Identified variations of this type include lengthening a measure that

- begins a hypermetric pair;
- effects tonal arrival or resolution of a dissonant chord;
- contains unexpected material, such as a highly chromatic chord in a diatonic context.

One intriguing feature of the non-parametric regression analysis is that the optimal Parzen window width σ leading to each tempo curve emerges out of the regression analysis through the process of cross-validation. The absolute value of

σ is usually in the range of 2–4 seconds (2.0591 secs for the curve of Fig. 1). It is an open question whether this value may hold some special significance, either in terms of tempo, or the structure of the piece, or even in terms of psychological properties of time perception and production.

4 The Hierarchy of Metric Deviations

The performed subdivision of each metric layer (e.g., quarter note) typically deviates from an even rendering of the next lowest layer (e.g., two equal eighth notes) as a function of time. This information can be represented in a graph such as that of Figures 3 and 4. The nature of such deviations varies with metric depth. They are often embedded in a small amount of random noise, which reflects limits in the perception and production of exact rhythmic ratios [15].

However, some systematic variations are noteworthy. For instance, a consistent lengthening of the metrically strongest half in a two-fold subdivision highlights its stronger metric position through agogic accent. This is in line with findings reported in many other approaches [7,8,9]. In our analysis, such specialized rules are generally arrived at by inspection, and are subsequently confirmed using standard statistical tests. The possibility of employing some machine-learning classifier to uncover such rules algorithmically, in a manner akin to [9], is currently under investigation.

It is perhaps most remarkable that, even though deviation from exact subdivision is free to vary on a point-by-point basis, the deviations observed in performance often vary smoothly over extended time spans, which typically corresponding to formal units such as phrases (see Figs 3 and 4). This suggests that manipulation of subdivisions is not always controlled on a pulse-by-pulse basis, which might impose excessive demands on real-time processing. Instead, it is shaped by broader gestures in a performer’s motor programs, coordinated so as to reinforce communication of musical structure. We would like to suggest that our multi-tiered analysis, which separates each layer of subdivision in the metric hierarchy, makes it easier for such patterns to be identified.

It should be added that the same non-parametric regression technique that is used to construct the tempo curve has been applied to subdivision timing data such as those of Figs 3 and 4, in order to extract the underlying envelope. Once that envelope is identified, one can seek rules that cause the subdivisions of *particular* pulses to deviate from the overall envelope. Such deviations can be typically attributed to the need to project some type of accent.

5 Conclusions and Future Directions

The present paper proposed a data analytic method that aims to uncover rules linking musical structure to specific expressive timing gestures in music performance. Several links were suggested between musical structure and expressive timing at one or several tiers in a hierarchy. The description of structure-to-timing associations remains to some extent qualitative at this stage. This could

perhaps be partly attributed to an inevitable element of unpredictability that may exist from performance to performance, even for the same player under different circumstances. However, given the present analysis, there are many ways to explore the possibility of precise quantitative relations between musical structure and expressive timing deviations.

For instance, correlations between structural features of the music and specific expressive deviations can be established by (i) annotating the score with a large number of potentially relevant features, some of them objectively identifiable (e.g., location of cadences), and some requiring annotations by independent musical experts; (ii) seeking correlations between the above features and expressive deviation gestures such as peaks in the tempo curve, lengthened measures, or lengthened beats. Such features can be tabulated in contingency tables, to which standard statistical tests can be applied.

Another analytical approach might involve modeling the exact shape of the tempo curve, seeking a quantitative predictive model of tempo fluctuations as a function of specific musical features. This would entail (i) quantification of all the possibly relevant features as continuous functions of time [16], and (ii) complex regression analysis to identify features that are the best predictors of the tempo curve. We are currently exploring certain multivariate time-series models that could lead to such quantitative relations. As for the expressive subdivisions within each layer of the metric hierarchy, they can be effectively modeled as they unfold in time using the technique of Hidden Markov Models.

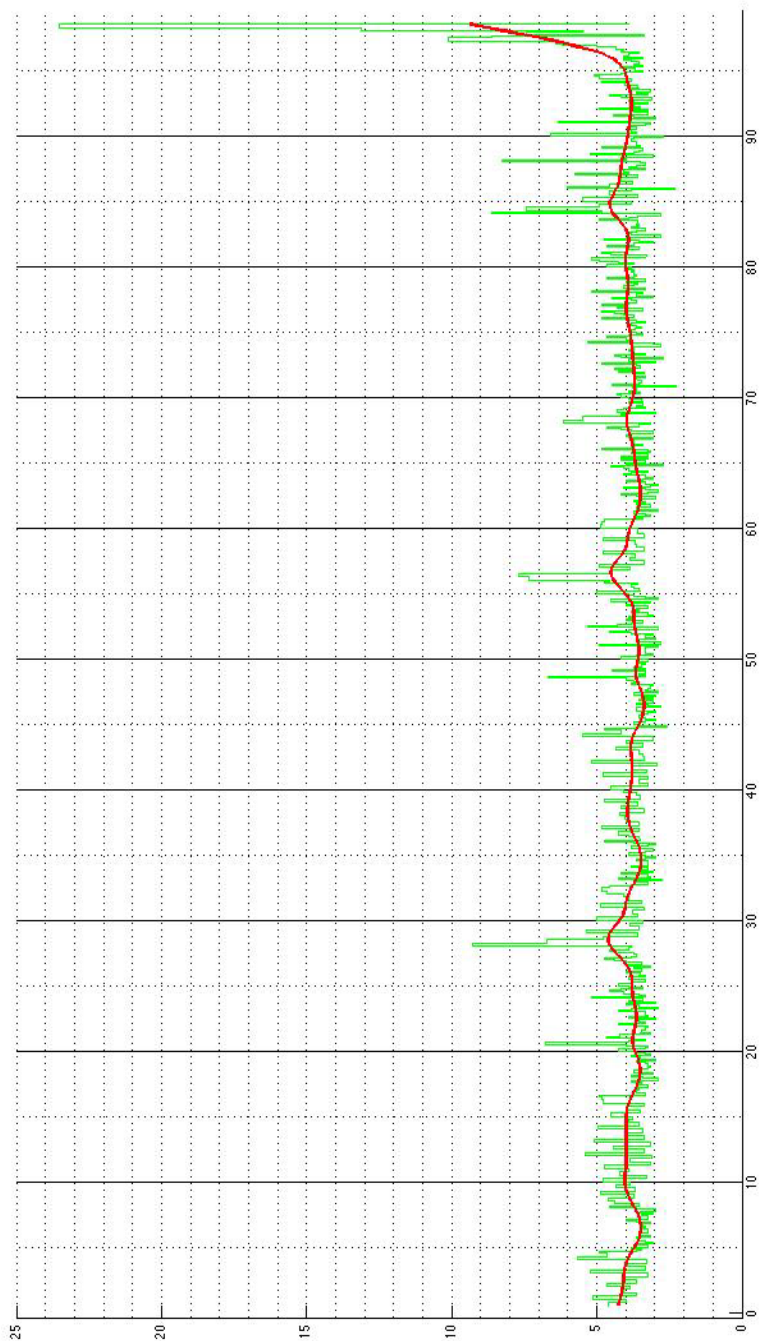


Fig. 1. Tempo curve for a recorded performance of Bach's F minor prelude, BWV 881 (WTC, Book 2), plotted against normalized inter-onset durations for each note in the piece. The x-axis represents position in the notated score using measure numbers (e.g., 2.5 is the middle of the second measure). The y-axis represents local tempo, measured by the duration of a whole note in seconds.

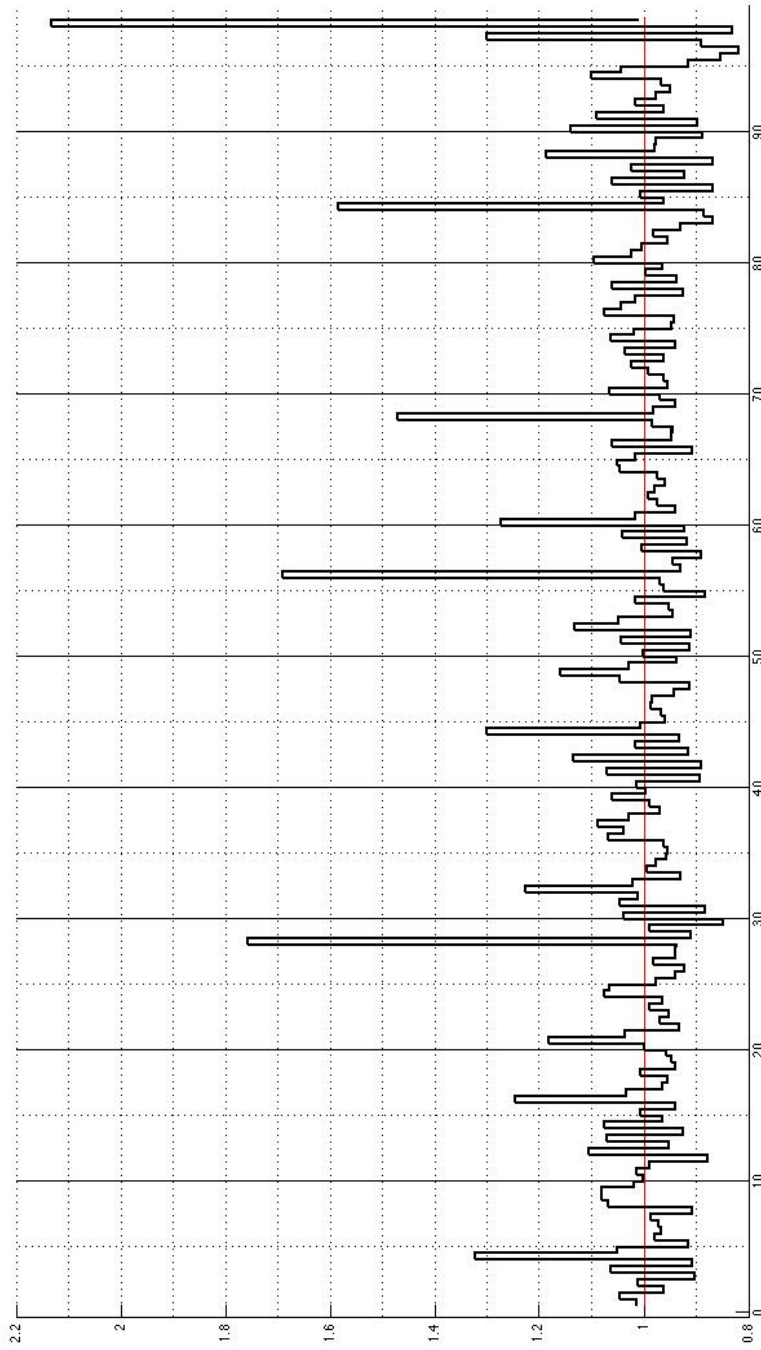


Fig. 2. Graph showing how individual measure durations deviate from the duration corresponding to performed tempo at each point in time. The x-axis represents position in the notated score using measure numbers (see Fig. 1). The y-axis represents individual measure durations in seconds. Therefore, a high spike represents a markedly lengthened measure. The graph comes from the same performance as that of Fig. 1.

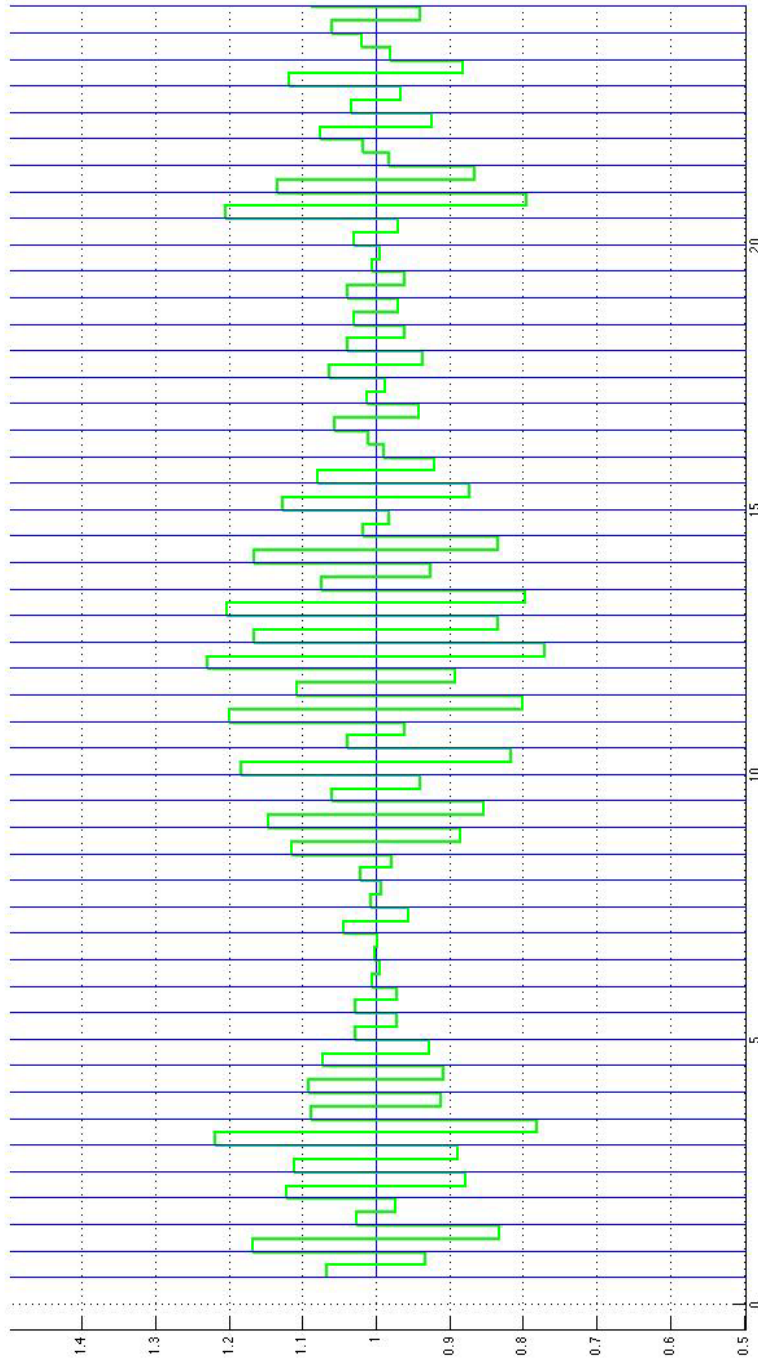


Fig. 3. Graph showing the performed subdivision of each quarter note into two eighth notes. As before, the x-axis represents position in the notated score using measure numbers (see Fig. 1). The y-axis represents the duration ratio of each subdivision. E.g. the first quarter note in the Figure is subdivided into two eighth-note time spans of duration ratio 1.06 : 0.94. The measurements are taken from the same performance as that of Figures 1–2. Figure 3 focuses on mm. 1–23. The complete piece is represented in Fig. 4.

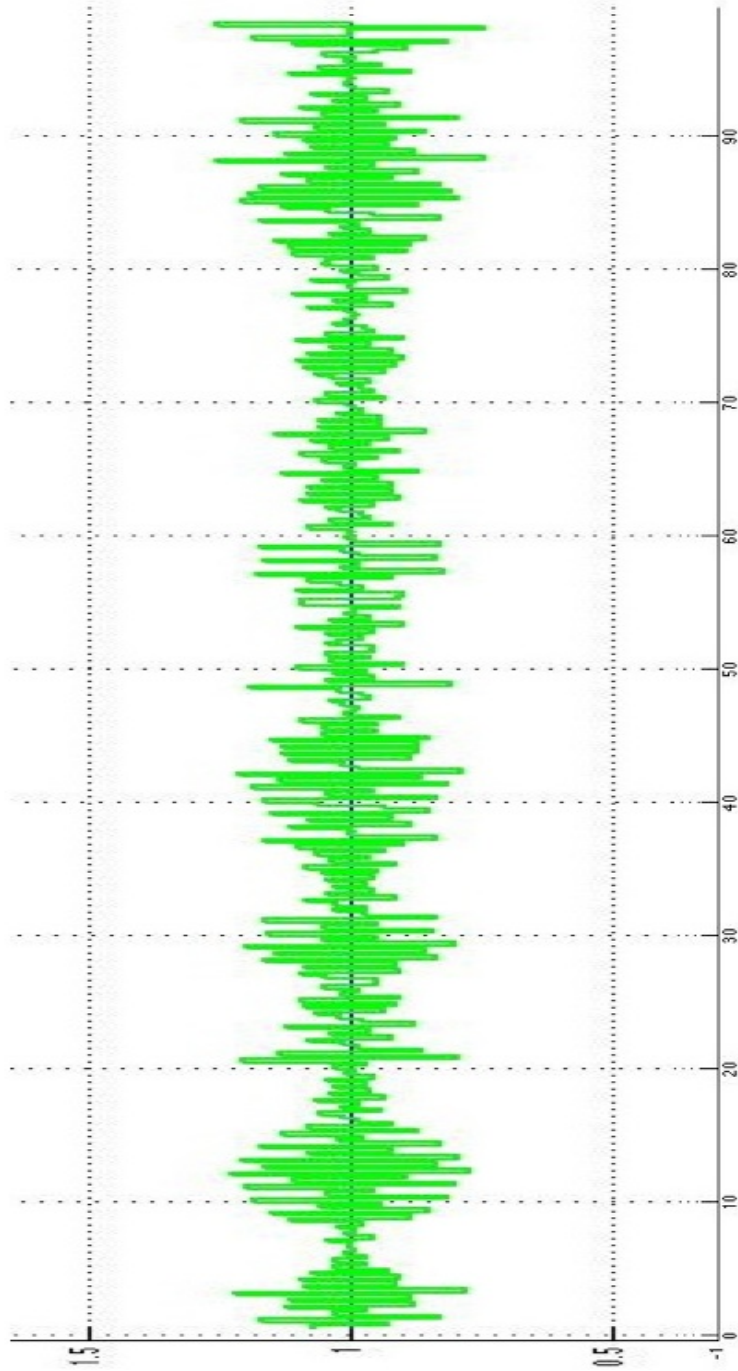


Fig. 4. The measurements presented in Fig. 3, now shown over a larger time scale that encompasses the entire piece. The axes carry the same meaning as those of Fig. 3. From this graph, one can clearly observe how the deviations from exact subdivisions vary smoothly in magnitude over time, as witnessed by the smooth envelope to the curve. These smooth variations are shaped by gestures (lumps in the envelope) corresponding to meaningful formal units such as phrases.

References

1. Gabrielsson, A.: Music Performance. In: Deutsch, D. (ed.) *The Psychology of Music*, 2nd edn. Academic Press, San Diego (1999)
2. Widmer, G., Goebel, W.: Computational Models of Expressive Music Performance: The State of the Art. *Journal of New Music Research* 33, 203–216 (2004)
3. Lerdahl, F., Jackendoff, R.S.: *A Generative Theory of Tonal Music*. MIT Press, Cambridge (1983)
4. London, J.: *Hearing in Time: Psychological Aspects of Musical Meter*. Oxford University Press, Oxford (2004)
5. Todd, N.P.M.: A Computational Model of Rubato. *Contemporary Music Review* 3, 69–88 (1989)
6. Honing, H.: Computational Modeling of Music Cognition: A Case Study on Model Selection. *Music Perception* 23, 365–376 (2006)
7. Clarke, E.F.: Generative Principles in Music Performance. In: Sloboda, J.A. (ed.) *Generative Processes in Music: The Psychology of Performance, Improvisation, and Composition*. Clarendon Press, Oxford (1988)
8. Friberg, A.: *A Quantitative Rule System for Musical Performance*. Doctoral dissertation, Royal Institute of Technology, Stockholm (1995)
9. Widmer, G.: Machine Discoveries: A Few Simple, Robust Local Expression Principles. *Journal of New Music Research* 31, 37–50 (2002)
10. Widmer, G., Tobudic, A.: Playing Mozart by Analogy: Learning Multi-Level Timing and Dynamics Strategies. *Journal of New Music Research* 32, 259–268 (2003)
11. Todd, N.P.M.: A Model of Expressive Timing in Tonal Music. *Music Perception* 3, 33–58 (1985)
12. Specht, D.F.: A General Regression Neural Network. *IEEE Transactions on Neural Networks* 2, 568–576 (1991)
13. Specht, D.F.: Probabilistic and General Regression Neural Networks. In: Chen, C.H. (ed.) *Fuzzy Logic and Neural Network Handbook*. McGraw-Hill, New York (1996)
14. Parzen, E.: On Estimation of a Probability Density Function and Mode. *Annals of Mathematical Statistics* 33, 1065–1076 (1962)
15. Clarke, E.F.: Rhythm and Timing in Music. In: Deutsch, D. (ed.) *The Psychology of Music*, 2nd edn. Academic Press, San Diego (1999)
16. Farbood, M.M.: *A Quantitative, Parametric Model of Musical Tension*. Doctoral dissertation. MIT Press, Cambridge (2006)