De- and recomposition of expression in music performance

Michiel Borkent

November 23, 2005

Final thesis INF
Student number: 9908986
E-mail: michielborkent@gmail.com

Supervisors:
dr ir Peter Desain - NICI, Radboud University Nijmegen
dr ir Jan Kuper, prof. dr ir Anton Nijholt and ir Eelco Herder -
University of Twente

Completed during January - November 2005

Art, engineering and science are - in that order - part of a continuum of finding truth in the world and about ourselves
– Richard P. Gabriel

Abstract

This thesis describes a method for de- and recomposition of expression in music performance. The focus in this thesis is on expressive timing: small deviations in onset timing that performers use to communicate structure or emotions in music. In the presented research expressive timing is identified as the result of interacting and overlapping structural units in the music. Examples of such units are (melodic) phrases, bars, and ritards. In contrast to approaches in the past, the presented model of expressive timing takes into account all structural units at once, without focussing in on one specific unit and leaving others out. The presented method of decomposition allows to decompose the expressive timing signal into structurally related profiles. For example, one profile accounts for the portion of ritard-related timing and another one accounts for the phrase-related timing. This makes a statistical analysis of performance in terms of structure possible. Moreover, it allows to edit expressive timing structure-wise. For example, the ritard-timing can be filtered out and phrase timing can be exaggerated, while keeping the rest of the timing intact. This idea of recomposition forms the second part of this thesis. Using the recomposition method new performances can be constructed from original ones. Easy application of creating stimuli for perceptual experiments, researching the relation between structure and timing, is now within reach. At the end of this thesis recommendations for further research are done, that will drive this research more towards an application inside recording studios, where performances could be altered more directly.

Preface

From January until this month, November in the year 2005 I completed another project at the Music Mind Machine group in Nijmegen, The Netherlands. This time in the light of my final thesis, which concludes the program of my studies at the University of Twente. I remember well that I enjoyed my first internship so much, that I was happy to do another project at MMM. Exploring the possibilities I wondered if it would it be like my internship, focussed on functional programming and Lisp, the language I learned to love in only a few months. Or would it be more in the direction of music cognition research? One thing was for sure: it was going to be a project that combined a few of my passions: music, mathematics and computer science. One other thing that seemed nice to me, was that my project would be creating or transforming music as output. Soon Peter suggested to read an article (Windsor et al., 2005) in which an analysis of music performance guided by a structural description of the music was explained. Peter told me that this article, of which he is one of the authors, would give rise to a program that could generate performances with edited expression. I was pretty much unaware of the scientific area that covers music performance, but I was ready to try to step into it. The project which is described in this thesis is the result of his suggestion and more than half a year's collaboration with Peter and others at MMM. Now having finished this project I can gladly say that it met prerequisites of combining my passions and that it helped me to discover many new things.

I would like to thank Peter once again for being a great hub into the world of computer music. Without him and the rest of his MMM crew I wouldn't have finished my engineering program in such a fascinating way.

A lot of other people encouraged and helped me throughout this project. This is the place to thank them. In no particular order thanks go out to: Barbara, Rudolf, Jan, Eelco, my parents, Luke, Edze, Ben, Alex, Yvonne, Makiko, people from De Verwondering, Haan, Esther, Dagmar, Renée, Rebecca, Bas, Irene and Bert and the people who I forgot or helped and encouraged me after this thesis was in print. Last but foremost I would like to give thanks to God. I have the impression He brought me to MMM and I have the impression He will take me to other interesting places in the future! - Michiel

Contents

Chapter 1

Introduction

In this thesis a method for de- and recomposition of expression in music performance is presented. This chapter gives an introduction to the research field concerned with expression in music performance and the position of the presented work in that field. It is recommended that the readers who are not known with certain terms and concepts used in this thesis, assist themselves by consulting the Glossary and index in the back of this thesis.

1.1 Setting

For many years music has formed an interesting domain for philosophy and science. From Pythagoras on man has tried to associate music with mathematic structure and theory. A wide area of music scientific research exists today.

Music performance

A subset of music research is concerned with music performance: the way a performer plays a musical piece, possibly based on a score, i.e. a symbolic representation of the musical piece. In this field the distinction between two time scales is often useful: the continuous time scale in which the onsets (starts) of notes from a performance can be displayed and perceived in time, and the rational time scale in which the intervals from the score can be displayed and are notated by the composer. To illustrate this, imagine that there are five stages and from the first to last stage the music undergoes a metamorphis. The first stage is creation: composition or improvisation. The second stage is, in case of written composition, the interpretation of written music. The third stage is the performance of music, the fourth the perception of the listener and possibly fifth, the remembrance of the listener. The transition between second to the third stage are of most interest to this thesis. The performer never exactly plays the notes' onsets and durations as prescribed by the score, the written music, originated in the first stage. Moreover, the performer often deviates from the prescribed timing on purpose to express emotions and emphasize structural entities in the music. This deviation is called *expressive timing*. Some examples: A buildup in tension can be performed as a gradual increase in tempo and a resolution of tension can be performed as a gradual decrease of tempo. It is often useful to compare performance and expressive timing with phenomena in natural language. Imagine someone reading aloud a written sentence: the structure of the sentence are the parts of which the sentence is build up. For example take the following sentence: "Every now and then, I see relations between different areas of science." It can be decomposed into two smaller parts, divided by the comma. When the person is speaking this sentence note the little break after the comma and note the even larger break after the dot. In music there are melodic phrases to be played by the performer. Often are these phrases built up by smaller parts that end in a ritardando (ritard from now on) and fermatas. A ritard is a gradual decrease in tempo and a fermata is a relatively long pause. A ritard can be compared with the comma in the sentence and the fermata can be compared with the dot. Many of these timing related structural entities, as they will be referred to, can be compared with counterparts from natural language.

Figure 1.1 illustrates the concept of expressive timing. The inter-onset intervals (IOIs for short) from a performance are compared with the prescribed inter-onset intervals from the corresponding score. IOI is the interval between two onsets (starts) of succeeding notes. The formula used to compare the performance IOIs with score IOIs is *inverse local tempo*. Loosely speaking, this formula says: the more time a performer takes between onsets of succeeding notes, the higher the inverse local tempo is (correspondingly, the lower the local tempo is). Since the durations in the score are static (they do not change after they have been decided), IOI and inverse local tempo are almost the same, except that when the term inverse local tempo is used, there is a distinct connection to an external norm: a prescription of what durations should be 'normally', usually the score[1]. For a more elaborate explanation of inverse local tempo, see the Glossary. The horizontal axis in the figure is the score time (in seconds) and the vertical axis indicates the inverse local tempo in the performance. Notice the peak around score time 25 s. that reflects a high inverse local tempo around the fermata in the score (measure 14, figure 1.2). The term fermata is explained some more in the Glossary.

Expression and structure

A part of music performance research is concerned with the relation between music performance and musical structure. Since the research of Seashore and

[1] In case of composition, the score can be used for this. In case of improvisation there can be made another norm, such as a quantized version of the improvisation.

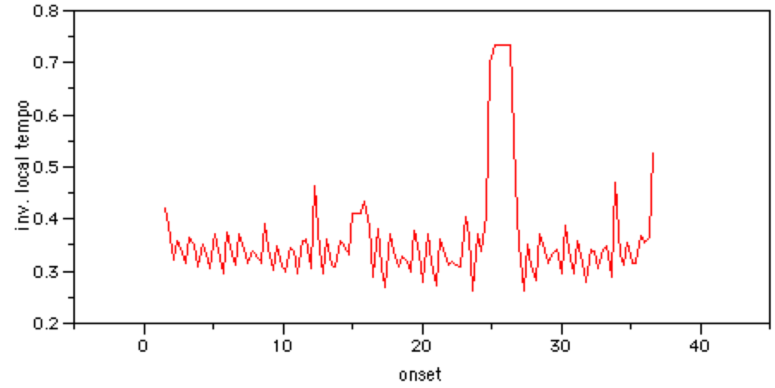


Figure 1.1: inverse local tempo diagram of a performance

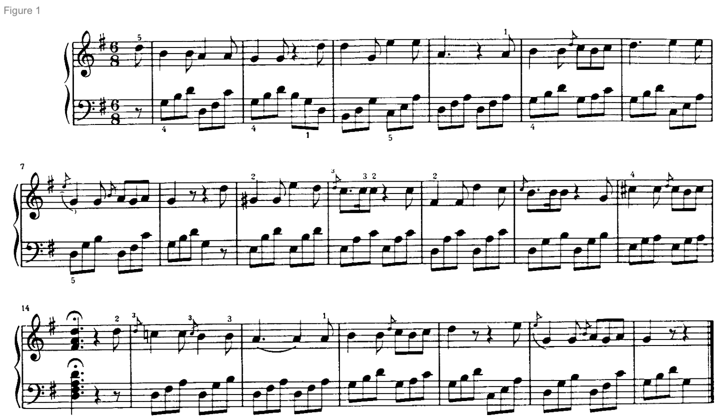


Figure 1.2: score representation of the music

colleagues (Seashore, 1938) there is scientific evidence that skilled performers use expression in structured and predictable ways that are related to the structure of the music. This gave rise to the idea that many aspects of expressive timing (and dynamics) can be predicted from an analysis of the structure of a piece of music, such as the fermata in the above score, and that such predictions can be formalized in a system of rules (Palmer, 1997). Many algorithmical approaches to this mapping of structure to performance have been developed. For example, Clynes (Clynes, 1983) predicts timing and dynamics from time-signature and composer characteristics, recursively subdividing time intervals multiplicatively at each metrical level[2]. Other approaches focus on other specific structural units such as leaps (jumps in pitch) or phrases. The notion of leap is explained further in the Glossary. A major benefit of such models is that they can be fitted to empirical data, yielding an estimate of their predictive power and optimal parameter values. However these models have helped in building and testing theoretical concepts about music performance, they are in general not very successful when fits to real performances are attempted. This might be caused by the fact that these models often focus on only one or a few, but not all structural units that contribute to the expressive timing signal. In the next subsection, in which we explain the aims of this research examples of these structural units are given.

Aims of presented research

This research is part of the performance-structure domain described hereabove. It takes a different approach to reveal the relation between structure and expression. First of all, all structural units that are assumed to contribute to the expressive timing pattern in a performance are considered at the same time. Secondly, this approach does not use a prior set of rules but uses a set of adjustable profiles by means of parameters, instead. Each profile corresponds with a structural unit in the sense that the expressive timing pattern can be thought of as the sum of all the individual profiles. The aims of this research can be summarized as follows: firstly, to decompose and analyze expressive timing in relation to the structure of the music and secondly, to reconstruct performances with adapted, yet musical meaningful expressive timing. This adapted timing is based on profiles derived from decomposition.

The title of this thesis might suggest that the current research has a more general focus than just expressive timing. The reason we chose this title is that the methods for de- and recomposition in fact can be applied to any expressive parameter, for example also to intensity (loudness, dynamics), although this thesis only treats expressive timing.

[2] Several metrical levels can be distinguished in a piece of music, all based on the meter, with the *tactus* being the most salient

Support of the research

We can ask ourselves the question: how useful is it to form theories and a science of this domain? Apart from the answer that everything suspected of rational (according to a system or logic) behavior is worth researching there are a few others reasons that are worth mentioning here.

First of all, if a full fledged science of music is desirable, music performance cannot be looked over. Where is the music anyway? In the score, in the minds of composers, performers and listeners, in the body movements of the performers or in the air waves? Does music exist without anyone hearing it? These questions can lead to a lot of discussion, but one thing is for sure: performance plays a significant role in this space of where the music might be.

Secondly, a scientific theory of performance in relation to structure allows to analyze performances (and hence performers) in a quantitative way, as an alternative of judging performers' capabilities and consistency in a qualitative way. Comparison of performances on different tempi and of different performers become possible in a more scientific manner.

Thirdly, a scientific model of performance and structure allows to de- and recompose expressive elements, as described in this thesis. This gives the possibility to alter performance expression in a musical meaningful way. This way of editing can be a valuable tool in recording studio's and in production of stimuli for perceptual experiments.

1.2 Structure in music

To describe in more detail what structural units in music can look like, an overview is given in this section. The model presented in this thesis attempts to analyze expressive timing in terms of structural units in the music, because, and this is an important assumption throughout this whole thesis, the structure affects the strategy a performer can take with regard to expressive timing. Note that in the figures used in this section consist of two parts: a part of a musical score with annotation in the form of shapes and below the musical score the names of the shape which are commented on in the text.

A useful notion in musical structure is phrase structure. A phrase can be defined as a subsequence of a melodic theme that forms an autonomous entity. These phrases can in turn be subdivided in sub-phrases and can form a bigger phrase together with other phrases. For example, in figure 1.3 3-phrases (phrases of three eighth notes length) combine together into a 12-phrase where in figure 1.4 four 12-phrases combine together as one 48-phrase and in figure 1.5 three 12-phrases combine together as one 36-phrase. This phenomenon has a counterpart in natural language: words that combine into phrases and sentences. A performer can emphasize smaller or larger phrase structure by, e.g., acceleration in the first part of a phrase and deceleration towards the end. Different classical styles of music are known for emphasis on either small or large phrase structure. For example, romantic music is known for the emphasis on longer phrases.

Another important structural aspect in music is the division into bars, indicating the metrical structure. Performers sometimes accelerate at the beginnings of bars and decelerate at the ends, consistently throughout the piece. The bar structure in a score is generally indicated by vertical dashes.

Furthermore there are some structural elements of a local nature, such as the leap and ritard in figure 1.4 and the chord-ritard (fermata) in figure 1.5. A ritard is a salient slowing down at the end of a long phrase. Again the analogy with natural language can be drawn: a pause at the end of a phrase (comparable with ritard) or sentence (comparable with chord-ritard).

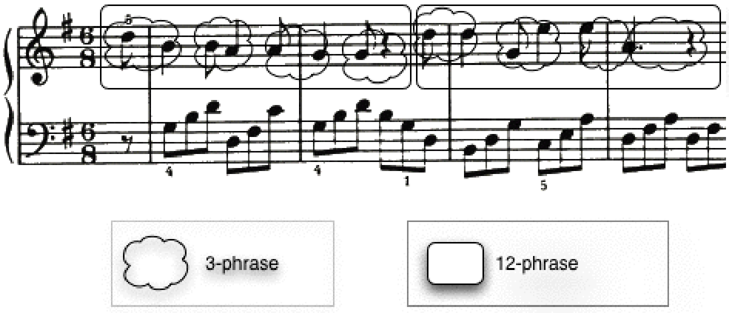


Figure 1.3: relatively small structural units

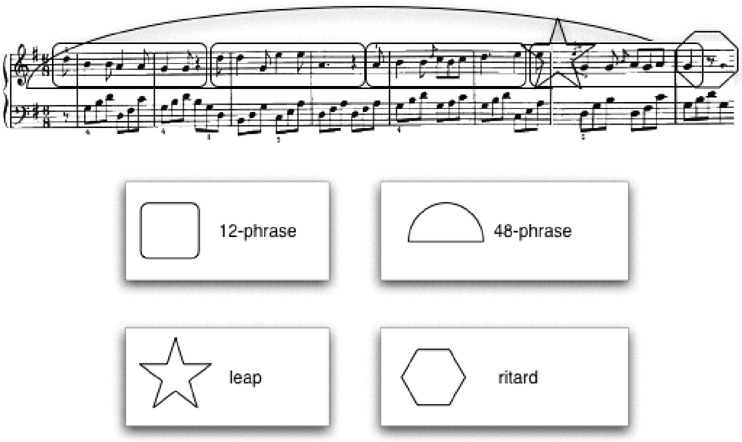


Figure 1.4: relatively bigger structural units

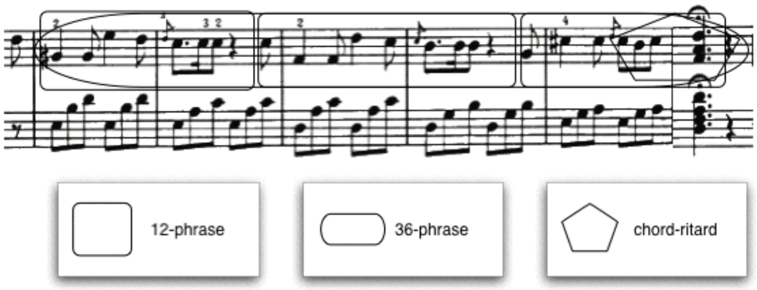


Figure 1.5: structural units

1.3 The performance and score data

The data used in this project is centered around the piece from Beethoven's six variations in G-major WoO 70 (1975) on the duet "Nel cor più non mi sento" from the opera "La Molinara" by Giovannna Paisiello (Figure 1.2). The theme is well suited for this kind of research since it has a lot of different and interestingly interconnecting structural units. For example, the metrical and phrase structures are out of phase by one eighth-note unit and both are regarded to play a role in generating expression.

The performances were originally recorded for Windor et al. (2001). The performer as was a professional pianist and instrumental professor at the Tilburg Conservatory in the Netherlands (age 26). The inter-onset timings of note onsets in the performances were captured using a Yamaha Disklavier MIDI grand piano. The performer had been given three weeks to prepare performances at 9 different tempi from the score in Figure 1.2. The nine different tempi were 50, 52, 55, 57, 60, 63, 67, 71, and 75 dotted quarter note beats per minute. This means that the duration between succeeding beats correspond with a dotted quarter note in the score. Of a dotted note the duration is one time and a half as long of an undotted note. These tempi were chosen because they span a reasonable range, yet are all within the bounds of technique and our musical taste. The pianist reported that although the more extreme tempi would not be his first choices, they were musically acceptable, especially after he had accustomed himself to them through practice. Of each tempo 5 repetitions were recorded.

A score with structural annotation (see Glossary and figure 1.6) of the piece was constructed in POCO (see Glossary and Honing (1990)). Although more than one structural description of the piece is possible, this one is chosen to be the most suited based on the intuition and experimentation of Luke Windsor and the first and third authors of (Windsor et al., 2005) with different structural descriptions and seeing how well the DISSECT method worked with them. See section 1.2 for more explanation on structural units.

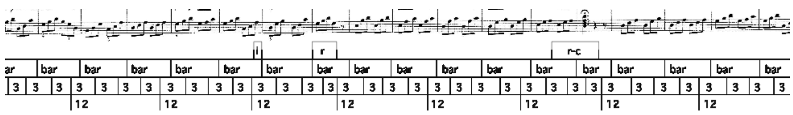


Figure 1.6: structural units

The score and performances were matched using score-performance-matching algorithms (Desain et al., 1997). In a matched score and performance pointers to corresponding notes are available. This makes the calculation of (inverse) local tempi possible by means of a simple algorithm that walks over the notes of the score and collects the onsets of corresponding performance notes.

Using POCO the onset times of all notes in the 45 performances were extracted, and the inverse local tempi 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 the analysis. A grace note is a musical ornament, that is, a short note played just before a main one.

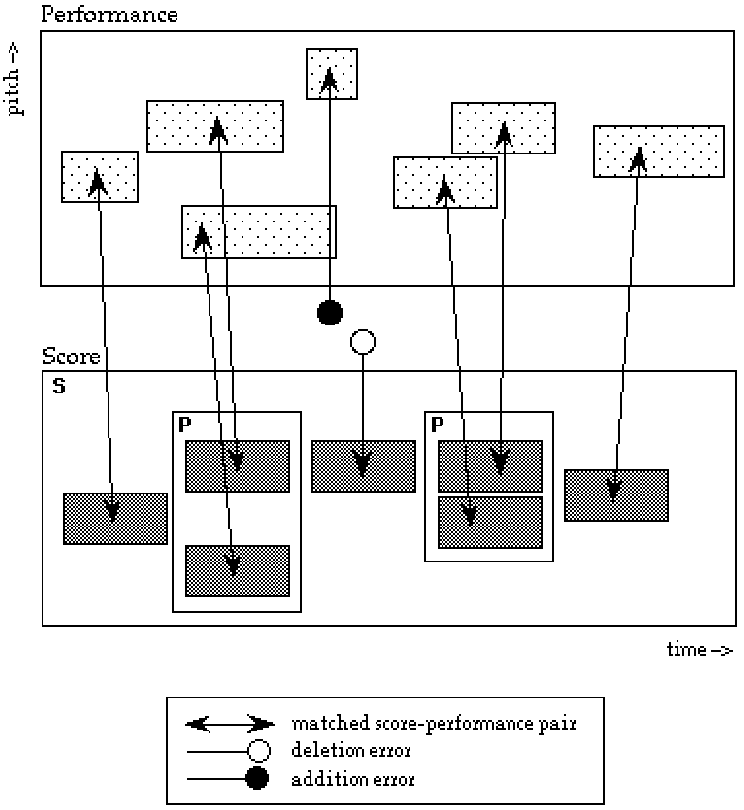


Figure 1.7: Matched score and performance

Chapter 2

Decomposition

2.1 Introduction

The previous chapter explains the setting of this research and the importance of an analysis of musical expression related to musical structure. It also explains about structural units that can be discovered in music that play a role with respect to expressive timing. Furthermore it gives a description of the used data. This chapter presents a method for decomposition of the expressive timing signal, extracted from this data (45 matched performances and structurally annotated score). Before the method, named DISSECT (with SECT standing for Structural Expression Component Theory) is explained, for didactical reasons a demonstration of a decomposition is given right away. The method is taken directly from Windsor et al. (2005). If the reader doesn't get it all right away it is recommend to just continue reading since this section only demonstrates briefly what the decomposition does.

At the bottom of Figure 2.1 the structural description of the score from 1.6 is repeated. A profile with a fixed length and with a fixed amount of free parameters is used to estimate the amount of timing that corresponds with each structural unit (see figure 2.2). The first assumption here is that for every structural unit of the same kind the profile is repeatedly the same. The assumption is that the expressive timing pattern is the sum of these repeated profiles. In short, the optimal parameters for each profile can be found as the solution of a multiple linear regression problem. In Figure 2.1 a decomposition of the expressive timing pattern into structurally related profiles is given, as a result of the DISSECT method, which is explained more detailed in the following section. Note that the lines in the repeated profiles might not look the same everywhere, while in fact the repetitions are the same. This is because the graph is aligned to the notes in the score.

Figure 5

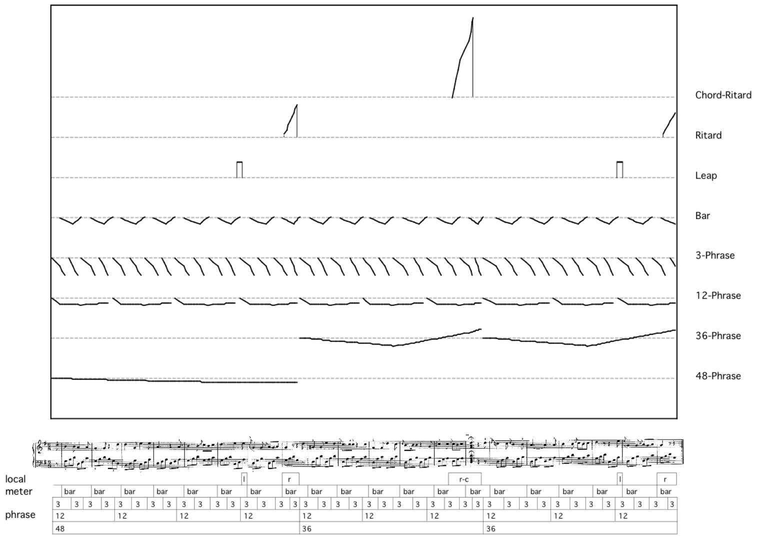
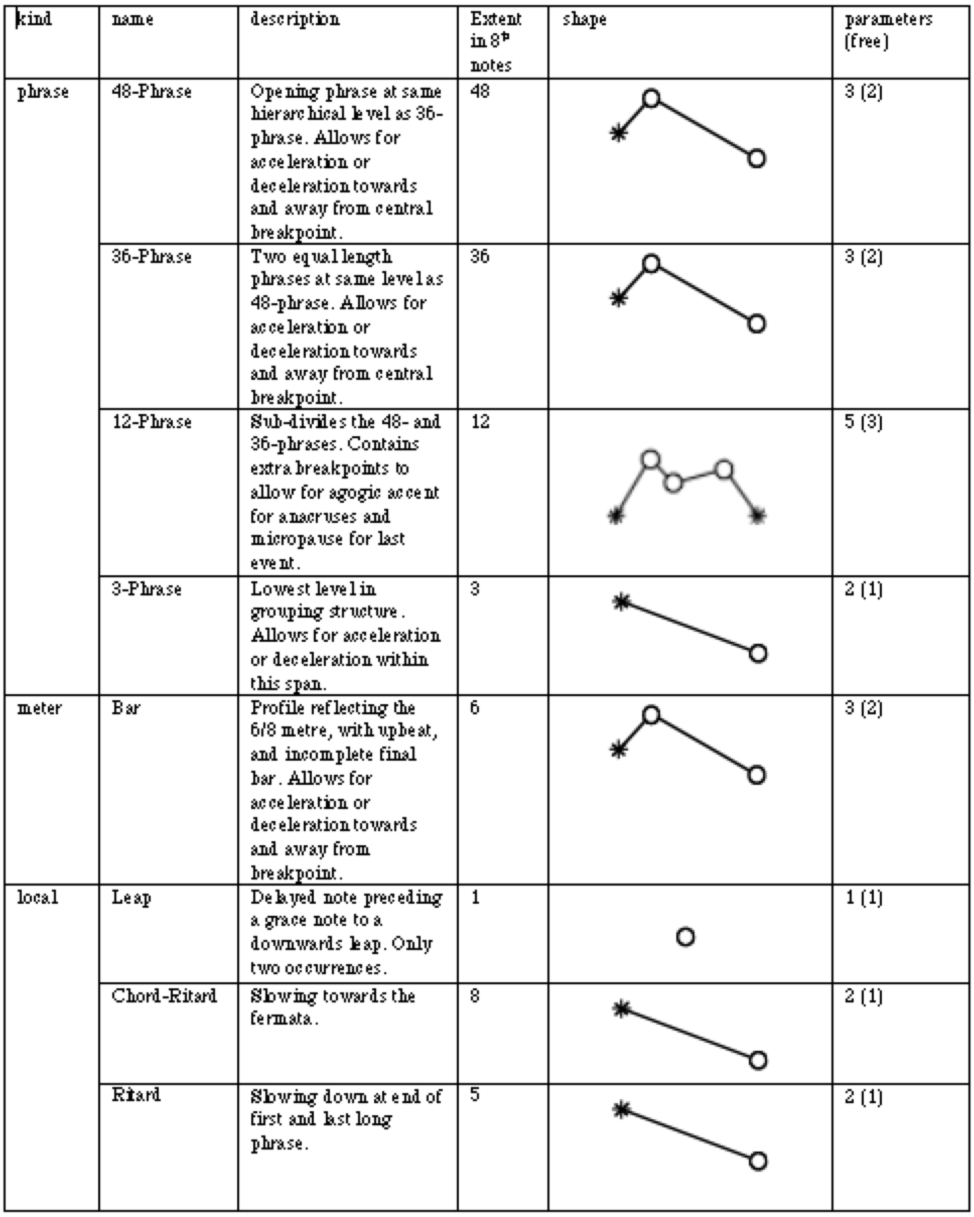


Figure 2.1: The expression signal decomposed into profiles, linked to the structural units - taken from Windsor et al. (2005)

2.2 Assumptions and procedure

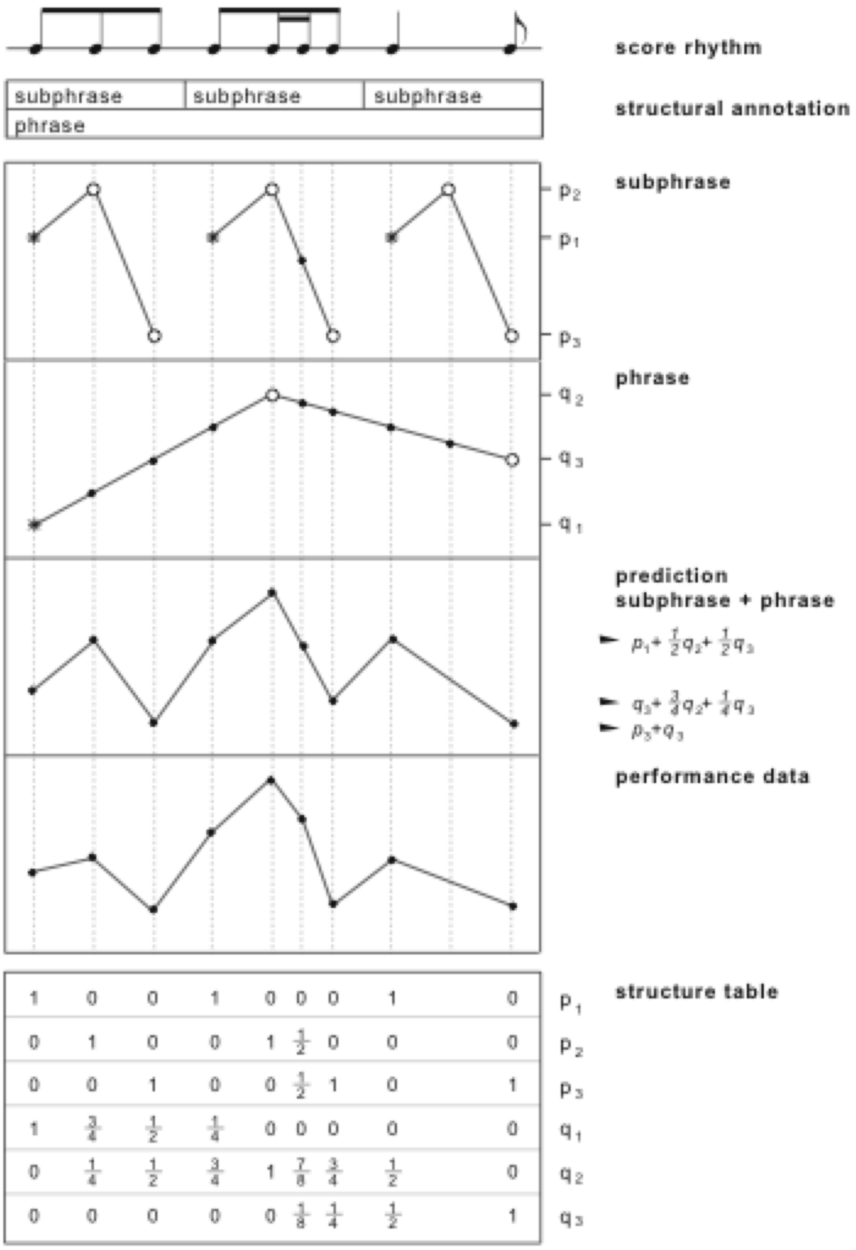
The method assumes that the expressive timing signal, expressed as inverse local tempo, is the sum of repeating and overlapping timing profiles, each one reflecting the expression of a distinct structural unit such as a (sub)phrase or a metrical level. The profiles are assumed to consist of line segments, with breakpoints specified at the first and last notes they span, and, if necessary, at one or more intermediate notes (usually a breakpoint in the middle suffices).

Figure 2.3 illustrates this using toy example. An imaginary score and structural description are given at the upper part. The profile associated with the sub-phrase is determined by the parameters p_1, p_2 and p_3 of which only the latter two are free parameters. A parameter to be free means its value can be adapted in contrast to a non-free parameter of which the value has already been determined. The profile associated with the bigger phrase, the combination of the three sub-phrases, is determined by the parameters q_1, q_2, and q_3 of which only the latter two are free. Note that the expressive timing signal is assumed to be the sum of the structurally related timing profiles, as stated before. Using this, the parameters that determine the profiles can be estimated by predicting the expressive timing signal as the sum of corresponding points in the profiles. In this sense, predicting means to find the



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)

Figure 2.2: Shape of profiles and their corresponding structural units - taken from Windsor et al. (2005)



1	0	0	1	0	0	0	1	0	p_1
0	1	0	0	1	$\frac{1}{2}$	0	0	0	p_2
0	0	1	0	0	$\frac{1}{2}$	1	0	1	p_3
1	$\frac{3}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	0	0	0	0	0	q_1
0	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1	$\frac{7}{8}$	$\frac{3}{4}$	$\frac{1}{2}$	0	q_2
0	0	0	0	0	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{2}$	1	q_3

Figure 2.3: Creation of structure table - taken from Windsor et al. (2005)

best parameters such that the differences between the predicted values (which are simply weighted sums of the parameters) and the observed data are minimized. For example, the inverse local tempo of the last note in last sub-phrase expressed as the sum of points from the profiles can be expresses as p_3+q_3. Why? Imagine p_1, p_2 and p_3 being points on a y-axis and q_1, q_2, and q_3 being points on another y-axis. Now look at the last note in the last sub-phrase and draw a vertical line cutting through the last breakpoints of the last sub-phrase and phrase. What values do both y-axes indicate at those points? Exactly p_3+q_3. This corresponds with the two ones in the column in the matrix on the line that was drawn from the note we were observing. Now for a more difficult example, the second sub-phrase predicted as a sum of points from the profiles can be expressed as $p_3 + \frac{3}{4}q_2 + \frac{1}{4}q_3$. Try do discover this yourself. In the same way the other points of the predicted inverse local tempo can be expressed in terms of sums of parameters, like captured in the structure table at the bottom of the figure. A similar matrix, say matrix A, is generated from the structurally annotated score in POCO. Now that the points of the overall profile can expressed as weighted sums of parameters, this overall profile can be fit to the performance data, that is, the inverse local tempi extracted from a performance. To fit profiles to the performance data is synonymous to predicting points the overall profile or predicting the optimal parameters. If the inverse local tempi to be predicted are expressed as vector x, with x_i being the inverse 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 Ap and observed x. Using the sum of square errors as a measure of difference this is a linear regression problem and can be solved as such: $p_{opt} = argmin_p||Ap||$, where Ap is the matrix product of A and vector p. The optimal parameter values in themselves have no particular meaning, but using them the timing profiles for each structural unit can be derived. For example, in figure 2.3 parameters p_1, p_2 and p_3 are used to predict the profile for the sub-phrase. The corresponding rows (the first three in this case because they correspond with p_1, p_2 and p_3) of the structure matrix and the optimal parameter values for these parameters (the optimal values for p_1, p_2 and p_3, say p_{opt_1}, p_{opt_2} and p_{opt_3}) can be used to construct the expressive timing pattern associated with the sub-phrase by matrix multiplication. In this example that would be

$$\begin{pmatrix} p_{opt_1} \\ p_{opt_2} \\ p_{opt_3} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & \frac{1}{2} & 1 & 0 & 1 \end{pmatrix}$$

More general, the timing profile for a certain structural unit can be derived by multiplication of the vector in which relevant optimal parameter values have their normal value and non-relevant optimal parameter values are set to zero, and matrix A. Selecting all optimal parameter values yields the predicted expressive timing pattern. One extra parameter is used to capture the global

BPM	correlation
50	0.92
52	0.96
55	0.97
57	0.97
60	0.96
63	0.96
67	0.97
71	0.97
75	0.89

Table 2.1: correlations between signal of model and original at different tempi

tempo of the piece (and a corresponding row in the structure matrix consisting of ones, also known as intercept) and lock some parameters to zero, the non-free parameters, in order to prevent the linear regression problem to become over-specified (matrix A becomes singular).

2.3 Analysis

Now that the expressive timing pattern can be captured in a model that assumes and decomposes this pattern as a sum of piecewise linear profiles linked to structure of the music, the decomposed signal can be used for analyses.

Goodness of fit

After application of DISSECT to a specific performance, it is interesting to see how well the predicted expressive timing pattern matches the original data. To test the goodness of fit 9 different performances were used, one for each tempo. Each performance is an average of the 5 repetitions at one tempo. The algorithm which makes an average performance from multiple performances calculates the average onset (and duration which is not important for this analysis but is sometimes needed for other purposed, such as playing it) of every corresponding note.

From the table can be concluded that the method works best with tempi 52 BPM (beats per minute) through 71 BPM and worst with the tempi at that are the most extreme (the slowest and the fastest). Nonetheless at these extreme tempi a reasonable model of the expressive timing patterns can be constructed that predicts the signal with a precision around .9 which is, to say the least, not bad.

Rep.	1	2	3	4	5
r^2	0.93	0.92	0.93	0.93	0.95
$avg.cr^2$	0.92	0.92	0.92	0.92	0.92

Table 2.2: cross validation of data and model from 5 performances at tempo 57 BPM

BPM	50	52	55	57	60	63	67	71	75
r^2	0.84	0.92	0.93	0.95	0.93	0.92	0.94	0.94	0.79
$avg.cr^2$	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.84

Table 2.3: cross validation of data and model from performances at different tempi

Cross-validation

It can be argued that the trade-off between parameters and data-points to be predicted can cause over-fitting: the model represents exactly the information available from the data, including noise. Using as much parameters as data-points results in a model that does not generalize over data, but is just another representation of the same data. In this section a cross-validation of the model from one performance to other performances is given to point out that our model generalizes well and is not the subject of over-fitting. The measure used is squared correlation which indicates how much of the data is explained by the model. r^2 indicates squared correlation and $avg.cr^2$ indicates the average of a serie cross-correlations. Cross-correlation is a correlation between the data of one performance with the model of another.

Table 2.2 gives an overview of cross validation between the 5 repetitions at tempo 57 BPM. The numbers in the first row represent the squared correlation of the data with the model of a repetition. The numbers in the second row represent the averages of the squared correlations between the model of a repetition with the data of all the other repetitions. (More formally, entry i of the second row is yielded as follows: $\frac{1}{4}\sum_{j\in\{1..5\}/\{i\}} r^2(model_i, data_j,)$, where $model_i$ is the predicted expressive timing signal of repetition i and $data_j$ is the original expressive timing signal of repetition j.) From this table can be concluded that the predicted data for one performance explains almost or equal as good for the data of other repetitions at the same tempo, hence it does not fit a lot of the noise that is present in an individual performance.

A similar table can be constructed to see how well models constructed of performance at one tempo explain data of performances at other tempi. It is reasonable to expect that the model from one tempi does not explain the data from a distant other tempo very well, since timing does not behave proportional over tempo (Desain and Honing, 1994). Table 2.3 presents such a table.

Structural unit	r^2	Prop. sd.	Stepwise r^2	Params. total (free)
chord ritard	0.786	0.823	0.786	2 (1)
3-phrase	0.08	0.284	0.456	2 (1)
12 phrase	0.055	0.08	0.181	5 (3)
ritard	0.006	0.177	0.173	2 (1)
bar	0.016	0.095	0.153	3 (2)
36-phrase	0.337	0.14	0.166	3 (2)
leap	0.023	0.087	0.107	1 (1)
48-phrase	0.012	0.068	0.075	3 (2)
Full model	0.948	0.982	-	21 (12)

Table 2.4: explained variance of individual timing profiles of an averaged performance at 57 BPM

From first row of the table can be seen that the squared correlation between model and data on the lowest and highest tempo is considerable lower. This observation matches table 2.2. The second row indicates that it is indeed the case that inter-tempo correlation is less high than inter-repetition correlation.

2.4 Importance of individual structural units

Another analysis that is made possible by the decomposition method is a measure of importance of individual timing profiles. As intuitively can be made clear that local accents can be more detailed on lower tempi and less important on higher tempi, it is now possible to verify this on a quantitative level. The measure of importance used is squared correlation: it tells how much variance in the expressive timing signal can be explained from an individual timing profile. Table 2.4 gives the result of an analysis conducted on an averaged performance at tempo 57 BPM.

The procedure that decides a table from an expressive timing signal and decomposed timing profiles, similar to table 2.4 is outlined column-wise.

The first column simply represents the names of the structural units involved in the analysis. The order of the names is decided as explained in the realization of the second and fourth column.

The order of the second and fourth column are yielded by searching for the profiles that in turn explain most of the expressive timing signal. The method of choosing this order is also known as step-wise linear regression. For those note familiar with it here follows a more formal description of how the order of these rows is chosen. Consider the original expressive timing signal e_{org} and an exact copy e which is changed among the procedure and the set S with tuples (s, t), s signifying the name of a structural unit and t its corresponding timing profile. For each tuple and correspondingly each timing profile, the squared correlation with the expressive timing signal is calculated: $r^2(e, t)$.

Structural unit	r^2	Prop. sd.	Stepwise r^2	Params. total (free)
chord-ritard	0.687	0.715	0.687	2 (1)
3-phrase	0.063	0.271	0.238	2 (1)
36-phrase	0.413	0.207	0.147	3 (2)
bar	0.029	0.147	0.103	3 (2)
ritard	0.006	0.199	0.094	2 (1)
48-phrase	0.012	0.088	0.04	3 (2)
12-phrase	0.005	0.054	0.027	5 (3)
leap	0.01	0.045	0.012	1 (0)
Full model	0.837	0.931	-	21 (12)

Table 2.5: explained variance of individual timing profiles of an averaged performance at 50 BPM

The tuple that scores highest is selected and its name s, squared correlation $r^2(e_{org}, t)$, and squared correlation $r^2(e, t)$ are appended to respectively the first, the second and the fourth column. Next the profile t is subtracted from e, yielding a residue of unexplained variance and the element (s, t) is removed from S. The procedure is repeated until S has become the empty set.

The third column, proportional standard deviation, is the standard deviation of the corresponding timing profile divided by the standard deviation of the expressive timing signal.

The fifth and last column shows how many parameters are involved with each structural unit in total, and how many parameters were free, in between parentheses.

From all individual performances can be constructed such tables, showing the order of importance of structural units. These tables can be used to observe individual performances and to compare multiple performances. As an example, it can be observed from table 2.4 that the chord-ritard is the most important, with respect to squared correlation that is, structural unit. As an example of inter-performance comparison, table 2.5 can be compared to table tbl:explainedvar-D. From this can be observed that the 36-phrase plays a more important role in the slower performance (the one at 50 BPM) than in the faster performance (the one at 57 BPM).

Chapter 3

Recomposition

With the expressive timing signal of a performance decomposed into timing profiles, associated with the structural units of the piece, a recomposition is possible. It is now possible to alter the individual timing profiles and generate a new expressive timing signal by adding them together. In this project the timing profiles can be manipulated by amplifying them with a weight (a real number). For example, figure 3.1 shows the original expressive timing signal of a piece and a recomposition of the decomposed signal with altered weights, according to table 3.1. Note that in the recomposition only the 3-phrase profile was used and was amplified by weight -3.3.

Amplification of a timing profile with weight 0 comes down to mechanical timing. For example, muting all timing profiles, or in other words, setting all weights to 0 renders a performance with mechanical timing. Amplification of a timing profiles corresponds with the performance timing of the timing profile. Setting all weights to 1 renders the original performance. Amplification with a negative weight comes down to turning it upside down and then amplificating it with the absolute value of that weight. Amplification with a weight that has an absolute value greater than 1 comes down to exaggeration of the timing profile, whereas an absolute value smaller than 1 comes down to understating.

Using this new signal a real performance can be reconstructed. This is elaborated on in section 3.2. First the recomposition of a new expressive timing signal is explained in more detail.

3.1 Limited effect of amplification

In recomposition of the expressive timing signal a distinction into three categories of timing profiles is made.

The first category consists of timing profiles that cover the whole piece. In our piece those are the bar and the 3-phrase profile. When these timing profiles would be amplified, normally the piece would become considerably shorter or longer, which is not desirable. For example, an exaggerated bar timing should

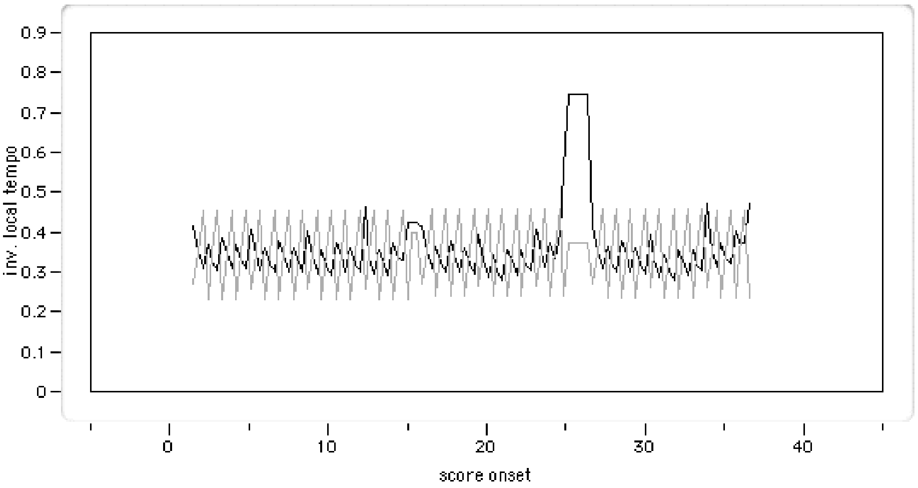


Figure 3.1: The original expressive timing signal (black) and a recomposition into a different expressive timing signal (grey)

Structural unit	weight
48-phrase	0
36-phrase	0
12-phrase	0
3-phrase	-3.3
bar	0
leap	0
ritard	0
chord-ritard	0

Table 3.1: Structural units and their contributions in a new expressive timing signal

not cause a bar to become longer in total - it should only affect the timing within the length of the bar. Therefore these profiles are manipulated before recomposition such that the average contribution to the expressive timing signal is 0. This is done by calculating the average contribution of the profiles and subtracting this average from every point in the profile.

The second category exists of timing profiles that are not intended to prolong or shorten the duration of the performance when amplified, like the profiles in the first category, but are not present everywhere in the piece. In our project these profiles are 48-phrase, 36-phrase and 12-phrase. Again the average of their contributions are calculated but only at places where they play a role in the piece the new expressive timing signal is lowered with this average (as opposite to the first category in which the profiles play a role just everywhere in the piece).

The third category forms the rest of the profiles which are simply amplified according to their weights, without further treatment. Especially with an amplified final ritard, it is obviously natural that the performance becomes longer. Ritards and the leap can be thought of in the same way: when they are amplified, the performance as a whole gets lengthier.

3.2 Reconstruction of performance

Now that a new expressive timing signal can be yielded from the timing profiles, this new signal can serve as a basis for a new performance. Note that the expressive timing signal itself is not yet a performance: it is only data about the inverse local tempi from some notes from the performance. Furthermore, information about intensity, duration and articulation is not available from this signal alone. Therefore, if it is our goal to generate performances that have the timing properties according to the new expressive timing signal, more information is needed. In this project the missing information is derived from the original performance, from which the original expressive timing signal was derived. The algorithm for reconstructing new performances is described below.

The performance is a set of notes that in turn are sets of note properties. Each note is determined by the following properties: key (pitch), onset time, duration, intensity and possibly some other properties not relevant to be mentioned here. For every note in the new performance the onset time has to be adjusted according to the inverse local tempo data. How that is done is described in the following paragraph. The procedure of decision of duration is described after it. Key and intensity are trivial cases: they are simply copied from the corresponding notes in the original performance.

Onsets

Only the onsets of the notes in the mixed expressive timing signal, melody notes and sometimes accompaniment notes, are known, as a result of recomposition of the timing profiles. The rest of the notes must get their onset information from a combination of the mixed expressive timing signal and their onsets in the original performance. There are three cases to consider when the onset of a note cannot be derived directly from the mixed expressive timing signal. Firstly, it is possible that that the score onset of a note under consideration is the same as another note that was included in the decomposition analysis. In that case the asynchrony in the original performance between these notes is calculated and used to calculate the new onset: the known onset of the note in the mixed expressive timing signal compensated with the asynchrony.

The second case is that a note is a grace note. The score onset of a grace note is equal to a main note (in the digital version of the score, not necessarily in the notated version) to which the grace note is an ornament. Similarly with the previous case, the asynchrony with this note, of which the onset is known, due to inclusiveness in the analysis, is calculated from the original performances and used to compensate its onset in the mixed performance.

The third case, the least occuring case, is that there is no note of which the score onset is equal and is included in the analysis. In this case the asynchrony with a neighboring note is calculated and used to calculate the new onset.

Durations

The durations in the new performance could be handled as trivial as keys and intensities. This is not the approach taken here, because of the fact that articulation in the new performance can be saliently different from that of the original if durations would be simply copied, since onsets can shift back and forth as a result of recomposition with altered weights. Therefore not the durations are copied from the original, but the articulation in terms of durations is imitated. To define articulation in terms of durations three cases are to be considered: legato, staccato and tenuto.

Legato

Consider two succeeding notes. They overlap if the duration of the first note, say d, is longer than the IOI of the two notes. This case is called legato. Consider the overlapping portion Δt, $d - IOI$. In the new performance, the duration of the first note is adapted such that the overlapping portion with the next note is exactly again Δt. Figures 3.2 and 3.3 illustrate this. In the original performance, the first note played starts at time $= 0s$ on key 60. After a while, a second note is played when the first note is still sounding. The overlapping portion Δt is used to calculate the duration of the first note in the mixed performance.

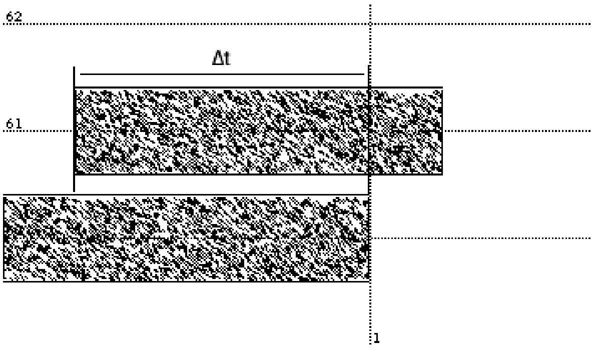


Figure 3.2: Notes played legato in an original performance

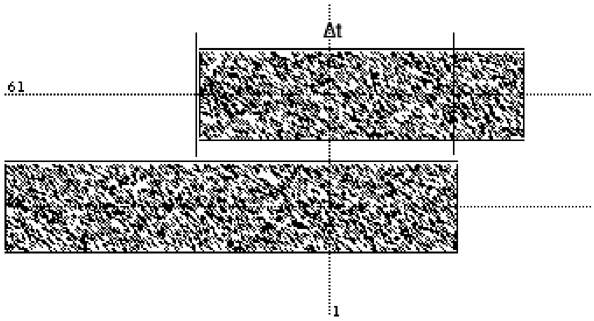


Figure 3.3: Notes played legato in a mixed performance

Staccato

The first case is when the duration of a note is shorter than a determined fixed real number. In this project this value was taken as .2 seconds. If this is the case, then it is decided that the duration of such a note should be maintained in the new performance.

Tenuto

The remaining case is called tenuto: the two succeeding notes do not overlap, but the first one is not short enough to be recognized as staccato. In this case the proportional duration, this is the fraction of the old duration d_{old} and IOI_{old}, the IOI of the two notes in the original performance, is maintained: the duration of the first note of the two in the new performance, d_{new}, is chosen so that the new proportional duration $\frac{d_{new}}{IOI_{new}}$ equals to $\frac{d_{old}}{IOI_{old}}$. This is illustrated by figures 3.4 and 3.5. The first note played has duration of $1s$, if followed by a pause of .5 seconds and succeeded with a note of duration $1s$. Note that the sound portion of the IOI of the first note is .66 (a real number is used instead of the idealized $\frac{2}{3}$ fraction to distinguish between performance and score time-scale). In the mixed performance the IOI is $3s$ and therefore the new duration of the first note is set to $2s$. Note that the original duration of the second note does not influence the new duration of the first note.

3.3 Demo

Now that the procedure for reconstruction of performances is explained, a demonstration is given of the whole procedure. Because this thesis cannot contain sound samples, a website with a demonstration was built in support of this section at `http://www.nici.ru.nl/mmm/demos.html`. Three performances in different styles were generated from one human-played performance. The expressive timing was decomposed according to the procedure described

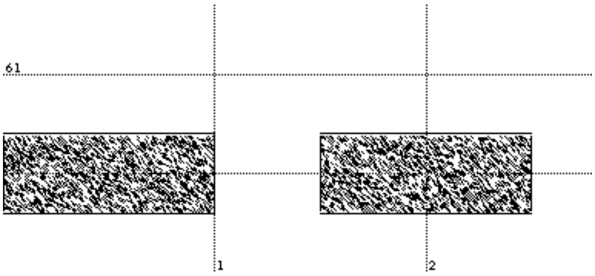


Figure 3.4: Notes played tenuto in an original performance

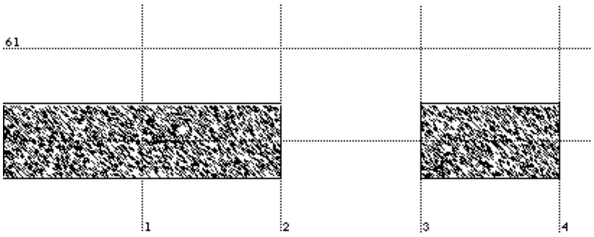


Figure 3.5: Notes played tenuto in a mixed performance

in chapter 2 and the resulting timing profiles were used to generate performances in renaissance, romantic, and ragtime style. Below the expressive timing diagrams of the mechanical, original and three mixed performances are given. All recompositions were generated by the expertise of Ben Turner, a music-cognition graduate from the Ohio State University. The values of the weights are based on the intuition complemented with the necessary tweaking of this student.

The mechanical version of the performance can be rendered by putting all weights to 0, see table 3.2. Because all the structure-related timing profiles are muted, the only timing profile that provides values for timing is the intercept profile, which captures global tempo. The global tempo is the same everywhere in the mechanical performance, as illustrated by figure 3.6.

In contrast with mechanical timing, table 3.3 shows the reconstruction of all expressive timing that was extracted from the original performance.

The first recomposition is labeled *romantic*. It features long, full phrases in the melody, evenly spaced eighth notes in the accompaniment, and an understated fermata. Looking closer at how this performance is yielded (see table 3.4), it becomes clear that only the 12-phrase timing profile was used and was amplified with factor 3. See table 3.4 for the weights used and figure 3.8 for an illustration of the recomposed expressive timing signal.

The second recomposition is called *renaissance*. It alters the Classical pattern in the accompaniment, so instead of short-long-short, it becomes long-short-short, somewhat more characteristic of Renaissance dance music. In

Structural unit	weight
48-phrase	0
36-phrase	0
12-phrase	0
3-phrase	0
bar	0
leap	0
ritard	0
chord-ritard	0

Table 3.2: Mechanical reconstruction

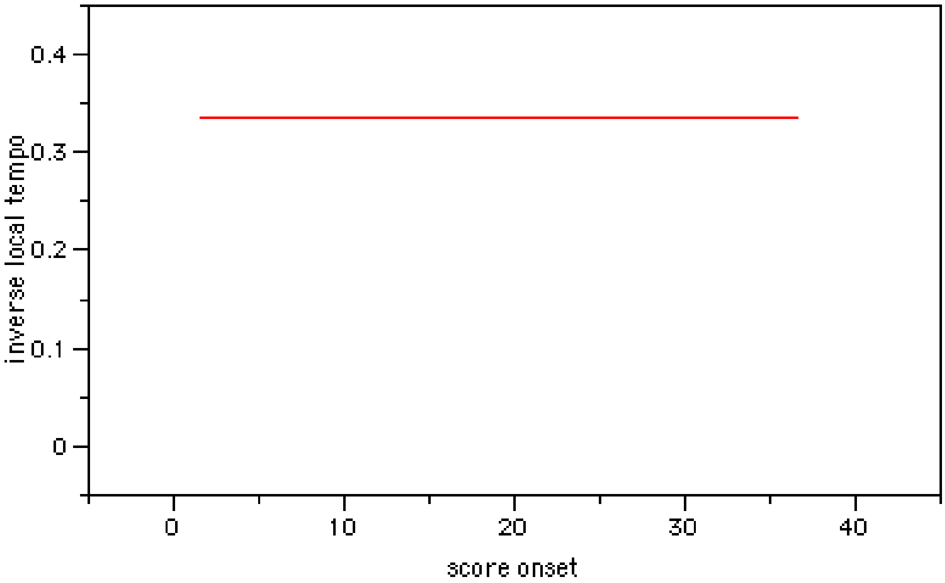


Figure 3.6: mechanical timing

Structural unit	weight
48-phrase	1
36-phrase	1
12-phrase	1
3-phrase	1
bar	1
leap	1
ritard	1
chord-ritard	1

Table 3.3: Full performance reconstruction

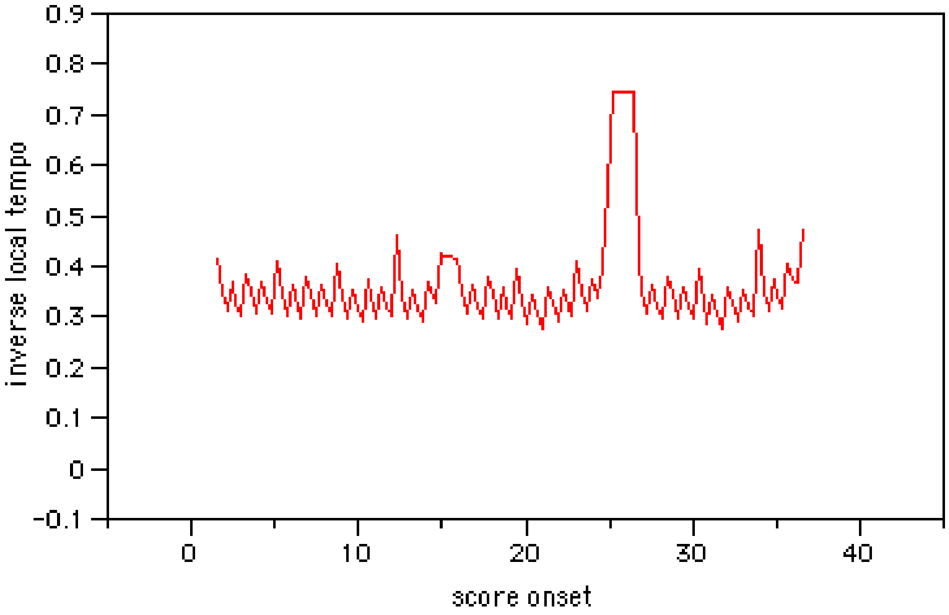


Figure 3.7: Reconstructed full performance timing

Structural unit	weight
48-phrase	0
36-phrase	0
12-phrase	3
3-phrase	0
bar	0
leap	0
ritard	0
chord-ritard	0

Table 3.4: Recomposition - romantic

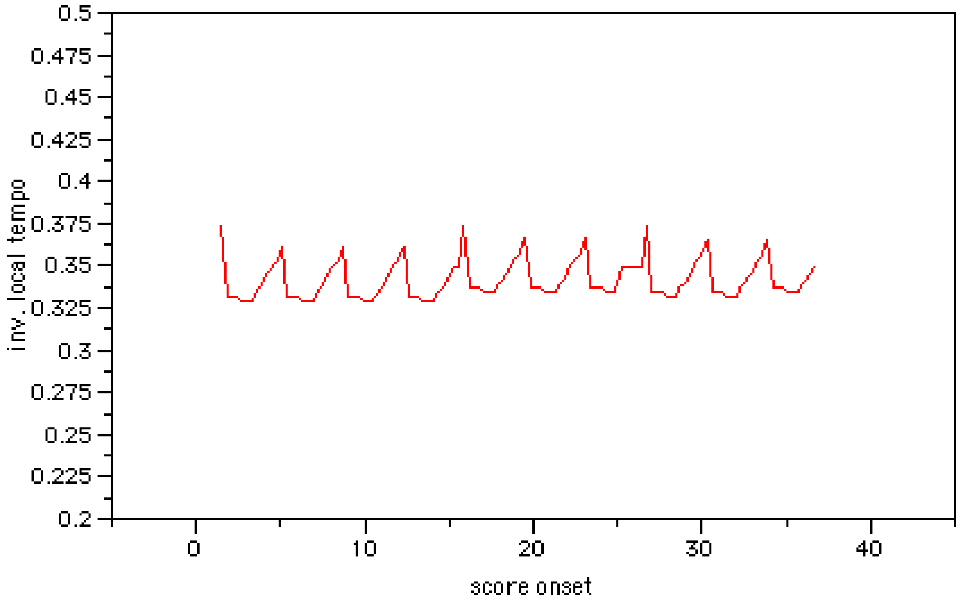


Figure 3.8: romantic

Structural unit	weight
48-phrase	0
36-phrase	0
12-phrase	0
3-phrase	-3.3
bar	0
leap	0
ritard	0
chord-ritard	0

Table 3.5: Recomposition - renaissance

order to maintain danceability, retards are minimized. As explained in the beginning of this chapter, amplification with a negative value turns the timing profile upside down. This is the case with the 3-phrase profile which is even more emphasized by amplification with a factor three. For this, see table 3.5. As with the romantic performance only one timing profile is sufficient to render a characteristic and realistic performance. See table 3.5 for the weights used and figure 3.9 for an illustration of the recomposed expressive timing signal.

The third and last recomposition is called *ragtime*. Admitted, this style is a bit far fetched to be associated with Beethoven and therefore turns this mix more into a toy example but is nevertheless interesting to show the flexible application of the recomposition method.The mix features a highly exagger-

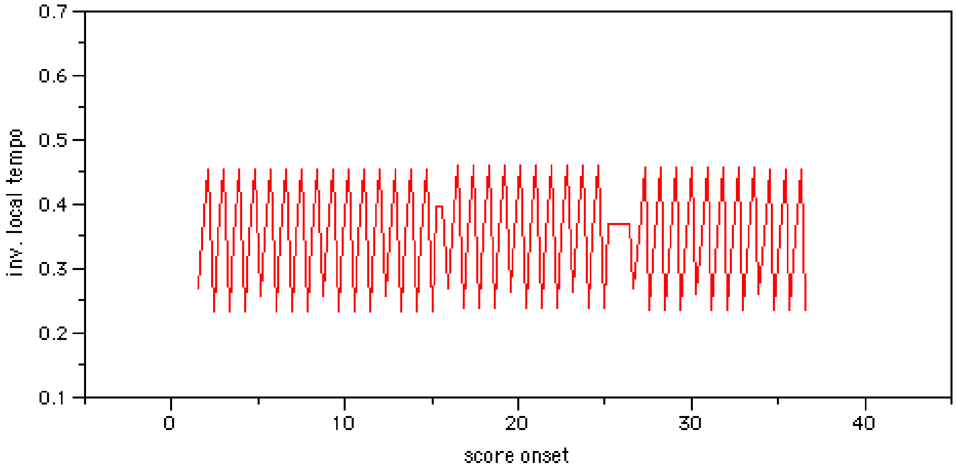


Figure 3.9: renaissance

Structural unit	weight
48-phrase	0.2
36-phrase	0.2
12-phrase	0
3-phrase	2.5
bar	0.2
leap	0
ritard	1
chord-ritard	1

Table 3.6: Recomposition - ragtime

ated eighth note figure which almost sounds like sixteenth-eighth-sixteenth in the accompaniment. Furthermore there is an exaggerated attention to the fermata. See table 3.6 for the weights used and figure 3.10 for an illustration of the recomposed expressive timing signal.

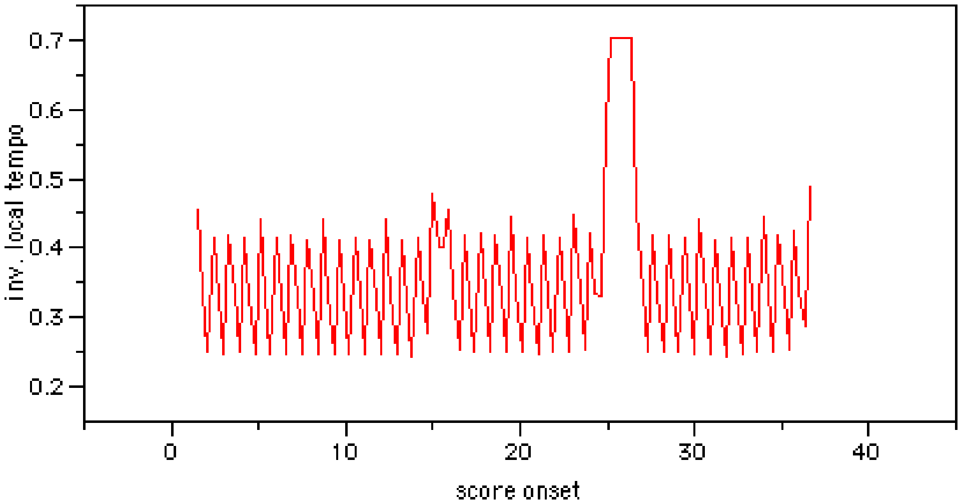


Figure 3.10: ragtime

Chapter 4

Design and Implementation

4.1 Organization

This chapter briefly report about the design and implementation of this project.

Programming tools

As mentioned, all modules involved in this project are extensions on the POCO environment, which is built on top of Mac Common Lisp 5.0. For more information on POCO see the Glossary. Common Lisp is a flexible multi-paradigm language which is very much suited for fast prototyping. The programming paradigms used in this project were mainly object-oriented, functional and imperative. In some cases the Common Lisp Object System provided helpful tools such as :before and :after methods which guided the structure of the programs towards a cleaner level.

Subdivision

Figure 4.1 illustrates the organization of this research project into modules. The arrows indicate the I/O between modules by means of file access. The parallelograms represent music files (scores and performances, possibly matched) which are either in the standard MIDI format (MID, 1997) or MTX (music text) format, which is similar to MIDI but with a few adaptations, such as readability (see the POCO entry in the Glossary for more information on MTX). A grey colored box with continued dash represents input or control information. A grey colored box with interrupted dash represents output. The hexagons represent arbitrary formatted files with e.g., structural information. The square-cornered squares represent tables and the rounded squares represent processes (Common Lisp programs) that manipulate the inputs and calculate the outputs.

Time

The implementation of the system took place in incremental steps during the months of February until October 2005, starting with the analysis method described in Windsor et al. (2005) and implemented as a POCO module. Furthermore the main part of the work consisted of the design and implementation of the design and implementation of a method to recompose and reconstruct mixed performances (the contributions ot During the project the existing module was split into separate programs to increase modularity and functionality. Here the Unix and Lisp motto was applied: KISS[1]. Functionality that was dependent on other software, like an external statistics package, was integrated and implemented within Lisp, such that the system became more autonomous. A POCO module was created from the SAPA (based on the book Percival and Walden (1993)) package (version 1.0), created by Donald B. Percival that handles the multiple linear regression work, involved in the decomposition analysis.

4.2 Control and interface

The interfaces of the decomposition and statistical components are pretty straightforward and more information about how to use them is included as documentation in the Common Lisp files. The interfaces of the recomposition and reconstruction components need some more explanation. As explained in chapter 3 the main control of the recomposition of timing consists of the weights that represent the amount of the contribution of each timing profile. Next to these weights there are a few other control parameters that can be used. For example, the profiles that should have in average have a zero contribution to the expressive timing pattern globally or locally (default set to bar, 3-phrase and 12-,36-, and 48-phrase respectively) can be controlled by separate parameters. This decision was made to leave the control of this decision in the hand of the user and with the eyes on the future, where possibly more locally and globally bounded structural units will be added. Other control options in the recomposition are: selection of source for intensity (defaults to the original performance) and source of articulation (defaults to original performance) which can for example also be set to 'full' which means that the durations of the notes will be as long as the IOI with the next note. This features might come in handy when distracting expressive parameters must be muted to be able to listen better to expressive onset time only. See section 5.2 to read about a proposal for real-time control.

[1] This acronym stands for: “Keep it simple stupid”.

4.3 Interaction

Now the interaction between the modules is discussed.

A structure table, as described in section 2.2, is created from a file with profile descriptions and an annotated score, which has information of where the structural units are located in the score.

A performance and annotated score are matched by a module here signified by *match performance and score* which was already available. The resulting matched performance-score serves two purposes in the system: extraction of local tempi and reconstruction of a mixed performance.

The expressive timing pattern, which can be derived by *extract inverse local tempi* and the structure table are used as input by the *fit procedure* which calculates the decomposed timing profiles, one for each structural unit and one extra representing global tempo.

Using a set of user-defined weights the timing profiles can be combined into an mixed expressive timing signal. Together with the matched performance-score *reconstruct performance* can, as its name says, reconstruct a performance that has the mixed expressive timing pattern and articulation of the original performance.

The timing profiles and the expressive timing pattern can also be used to do statistical analysis, as described in section 2.3.

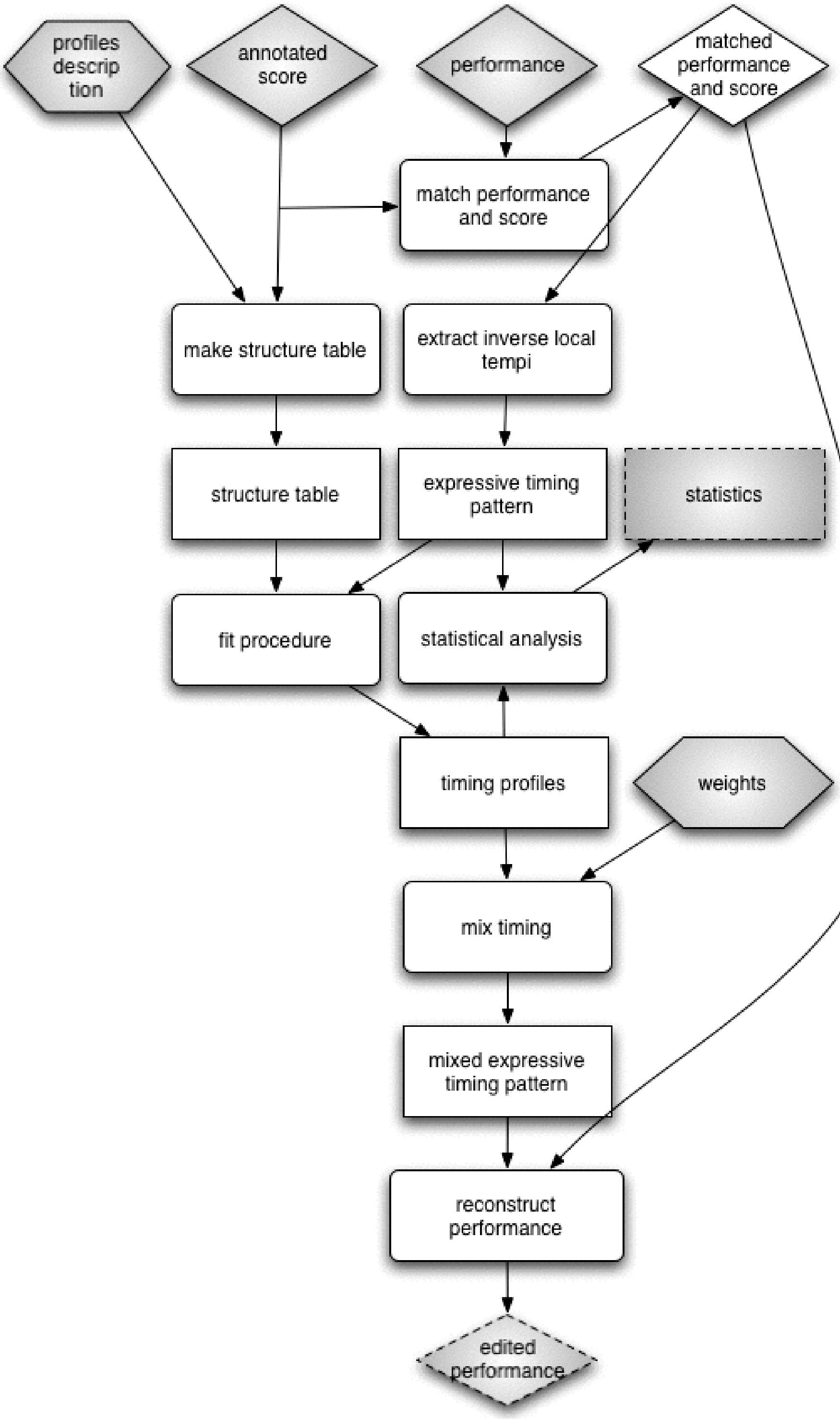


Figure 4.1: Diagram of modules involved in project

Chapter 5

Conclusions

This chapter gives some concluding words on the project and recommends points for further research.

5.1 Concluding remarks

In this thesis a method for decomposition of expressive elements in music performance was presented. This method makes a statistical analysis of music performance in terms of musical structure possible, as an addition to more subjective analyses of performance. Furthermore the method makes a recomposition method possible, as it was developed during this project. The combination of these two methods makes editing of structure related expressive timing possible. A demonstration was given to illustrate one of the interesting possibilities of the system: creating performances in seemingly different styles, based on one original performance.

One application of the recomposition method can be thought of in studios where performances sometimes need some alteration. Our system forms an alternative of letting a musician play his performance all over again.

Another application of the recomposition method is the production of stimuli for perceptual experiments. It can be investigated when and what kind of differences in emphasis on musical structure are most salient.

The system described in this thesis forms a step forward in expression editing. Some components of the system can be improved or automatized. In this light proposals for further research are listed in the following sections.

5.2 Further research

Automatic detection of structure

In this project structural components were recognized by a human expert who annotated the score with the corresponding markers. This task should be done for any piece of music involved in the decomposition process and takes up a lot of expertise, time and energy. It would be very efficient if this task could be performed by machine instead of human being. Moreover it would be one step towards real-time expression editing (also see 5.2). Much promise of the automatic detection of structure lies in the work of Lehrdahl and Jackendoff titled “A generative theory of tonal music” (Lehrdahl and Jackendoff, 1983). Although the ideas in this book do not give a mathematically detailed and closed formalism, they can form the basis for a true machine-conducted structural analysis. In fact, this is the approach taken in Hamanaka et al. (2005). Their approach is to solve the problem of unordered and sometimes ambiguous sets of rules by adjustable parameters, enabling to assign priority to the rules. Such a system would very much be suited to integrate with the one described in this thesis as illustrated in figure 5.1. The component outside the cloud is already implemented and forms the point of connection with the components within the cloud, which remain to be implemented. The cloud signifies the dreamy and idealistic atmosphere around the components which do not exist or are not integrated in the system yet, but hopefully one day will be.

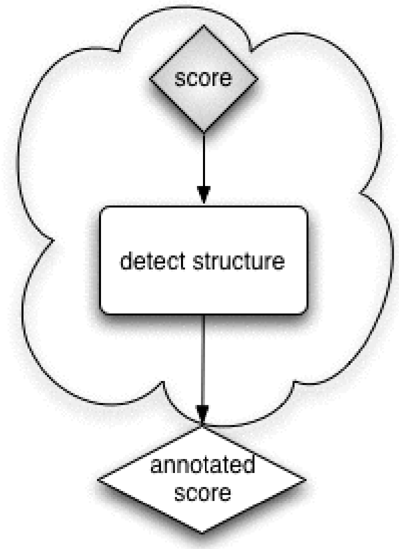


Figure 5.1: Integration of automatic structure detection

Quantization

The second suggestion for further research is integration of a quantization component. Quantization can be seen as the art of reconstructing the score from a performance. In other words, given a performance, find the most probable score used by the performer. Much research in the area of quantization has been conducted and good results have come into existence that could be used to be integrated here. In Desain and Honing (1992) a good deal of different approaches on the quantization problem are given. Suppose that a quantization component were to be developed. Figure 5.2 illustrates how it would integrate into the system, in combination with the proposal of automatic structural detection. The implementation of both components would drastically simplify the input of the system: a performance, instead of an annotated score and performance. Furthermore the score and performance do not need to be matched by an external matching component since the quantizer 'knows' what score note corresponds with the specific performance note; it generated the score note from the performance note itself. The arrows to the matched score and performance, which is seen as one object in this diagram (a bipartite interconnected graph of note objects, one component belonging to the score and the other to the performance) go from the quantizer (and hence matcher) and from the detection of structure. The order in which the matched score-performance is built up is not really focus here, but it could be like this. First the quantizer generates a matched score-performance which is the matched score-performance object but without structural annotation. This information also goes into the structure detecter and undresses the object until it has only the score. Then it does its works as it would do normally using the score information. After this has been completed the matched performance-score and annotated score are merged together again, resulting in a matched performance-score with structural annotation. Of course there are others ways to implement this system. That is the reason the diagram leaves the exact procedure obscured for now.

Having integrated a structure detecter and quantizer component, this would be the kind of system that to be used in studio situations. One other point of suggestion leads to this road, namely real-time control of the control weights.

Real time control of parameters

In many areas, whether the application of the machine one is controlling is in music or other areas like microwaves or even more evident cases like cars, the ability to control real-time is very desirable. The recomposition component of this project can be controlled by a set of weights, each weight controlling the amount of contribution of each timing profile. For example, if one want to filter out the ritard tempo delays, the according weight can be set to 0. This project is a first approach to this method of recombining timing profiles. The control of the weights is implemented as an offline solution. This means that one has to set the weights, run the program, wait and hear the results. Real-time control of these weights is not hard to imagine and has already been simulated by means of using time functions instead of real numbers as

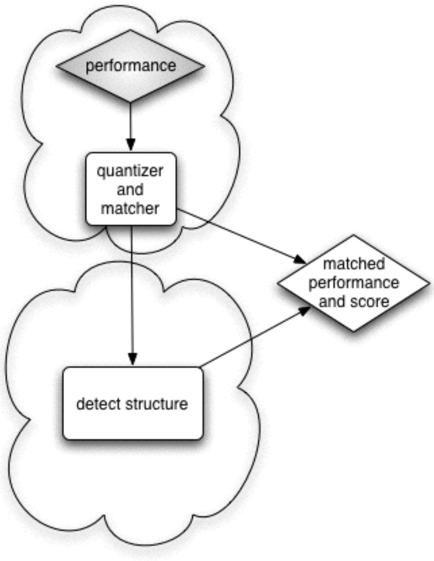


Figure 5.2: Integration of a quantizer

weights. The problem one has to face when designing and implementing real-time control is: when are the weights updates (read from the interface) but more importantly, when do they have effect on the generated music. Does the system allow the bar timing to change inside bars or only when a new bar begins? Because this is not the place to give a full decisive design on these issues, they are left to be picked up later.

Integration in webinterface

All the POCO modules can be used from the web, using POCO-web. In contrast with POCO, this web-interface provides an easy and stable interface to world users without the need to know about the modules and all kinds of other details that have to do with implementation such as the programming language Common Lisp. More information about POCO-web can be found at `http://www.nici.ru.nl/mmm/programs.html`. Integration of the modules involved in this project would allow to de- and recompose performances over the web. More information about POCO can be found in the Glossary.

Performance evolution

With the decomposition method it is possible to subdue performances to statistical analysis in terms of their structural properties. This can be used to characterize music from different periods. For example, the question can be raised whether the chord ritard was more or less prominent in older performances than in newer performances or whether the same performer played with different expressive timing in different stages in his or her musical career.

Integration with composition software

The timing profiles from a large collection of performances could be collected and used to predict expressive timing of new compositions. Application in composition software of this idea is interesting, because composers could directly hear their composition come alive, as expressive timing certainly adds more to the music since composition is only the first stage of music to enter the world.

Glossary

This chapter provides extra information for those of you who are either from outside music or science and like to have more information on certain definitions and subjects. It can be used as a reference in support of the rest of this thesis.

Computer music related

DISSECT is a verb that is almost synonymous to 'to decompose'. It is also the name of a method for decomposition of expressive timing in music performance, introduced by Windsor et al. (2005) with SECT standing for Structural Expression Component Theory which is described in chapter 2 of this thesis. The method uses an annotated score. Annotated here means that the score contains markers of where structural units begin and end. Structural units are parts of the score that bear a special relation with the expressive timing a musician uses in a performed piece of music. Examples of such structural units are bars, phrases and ritards. For more on this, read section 1.2. For more on the decomposition method, read sections 2.1 and 2.2.

MIDI stands for musical instrument digital interface and is a common and standardized protocol for communicating control information for electronic music instruments. In contrast to techniques which capture the sound of an instrument and play it back by regenerating wave forms from a stored file, MIDI only uses messages to control an instrument and does not contain any sound wave information itself. Examples of MIDI messages are: at time $0s$ press key number 60 (the middle C on a piano clavier), with full velocity (as hard as possible) and at time $2s$ release key number 60. These two messages can be put in a sequence and can be given to a sequencer, a device which sends the messages one by one at the appropriate time to a musical instrument that can interpret these MIDI messages. Such an instrument can be, for example, a sound card with wave tables, for example for every key number one wave file to be played or simply a digital piano.

POCO is a workbench for musical research, mainly conducted at the Music Mind Machine laboratory Nijmegen, The Netherlands, but which is also used through the web by many users worldwide. The workbench consists of various modules of different categories and work on performance and score files in the form of MIDI or its readable counterpart MTX. An example of such a category is conversion, for example to convert MIDI to MTX or convert MTX files with onset data to MTX files with IOI data. Another example is an analysis, which converts a performance to a table with information about the performance. Generally speaking the modules in POCO are centered around performance research, such as the research presented in this thesis. POCO is implemented in MCL (Mac Common Lisp). Non-Lisp users can benefit from POCO by using it through its webinterface POCO-web (see subsection Integration in webinterface in chapter 5.

Mathematics and computer science related

Common Lisp is the standardized. well documented and most used Lisp dialect. In the year 1958 John McCarthy invented Lisp, of which the name is derived from LISt Processing. It is one of the oldest programming language still in use and according to its users this is because of the language's flexibleness. Paul Graham describes in his book On Lisp[1], that, in contrast with the situation that a programming problem has to be translated to a programming language, Lisp can be used to grow a language towards the problem. Lisp is characterized as a functional programming language by many, but incorporates a lot of other programming paradigms as well, like imperative and object oriented. Moreover Common Lisp has the Common Lisp Object System which features the normal features one would expect from an object oriented language, like multiple inheritance, type dispatching and wrap-around methods, but also has the meta-object-protocol which allows the user to have full control over many things like in which way specific classes should inherit from each. Nowadays Common Lisp is used in software like Emacs, AutoCAD, POCO and OpenMusic and in many artificial intelligence research (see for example P. (1992)). A good reason to learn Lisp, even when you are not going to use it, according to hacker Eric S. Raymond is that Lisp will give you a feeling of deep enlightenment, once you get it. This experience will help you be a better programmer for the rest of your life, even when you hardly use Lisp itself. Richard P. Gabriel says about this programming language:"Lisp is the language of loveliness. With it a great programmer can make a beautiful, operating thing, a thing organically created and formed through the interaction of a programmer / artist and a medium of expression that happens to execute on a computer".

Multiple linear regression is a technique from the domain of statistics. Using a measure of error, usually the sum of squared differences, it is a minimization problem that finds the optimal or expected values of free variables given the values of other variables. In simple two dimensional form this can

[1] Free for download at his website `http://www.paulgraham.com`

be expressed as follows $y = \alpha + \beta x$. The variable of interest, y, is conventionally called the 'dependent variable' (also called 'input' or 'exogenous' variable) and x the independent variable. In the case that x and y are vectors, one speaks of multiple regression. For more information, see for example `http://en.wikipedia.org/wiki/Linear_regression`.

Music related

Articulation is a term used to describe the transition or continuity between notes. In this thesis three different terms are used to describe articulation, in terms of IOIs and durations: legato, staccato and tenuto. See subsection Durations in section 3.2 for a more elaborate explanation.
Dynamics is is term to describe changes in loudness, volume or sound intensity. In some areas people use the term loudness when the context is about human perception (humans perceive changes in sound with a logarithmical correlation with intensity), and intensity when the context is about the energy of the sound and objective sound measurements. A term that is used in the MIDI protocol to express dynamics is velocity: the force with which a key is pressed on the piano clavier. In music performance dynamics are an important tool for the musician to express changes in mood, locally and globally. For example, a build up in tension can be a gradual increase of intensity.
Fermata (or hold) is an element of musical notation indicating that the note should be sustained for longer than its note value would indicate. Exactly how much longer is to be decided by the performer, but twice as long is not unusual. See figure 5.3.

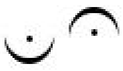


Figure 5.3: Fermata marks

Grace note is a common term for a phenomenon of music notation used to denote several kinds of musical ornaments. When occurring by itself, a single grace note normally indicates the intention of either an appoggiatura (from the Italian word 'to lean upon') or an acciaccatura (from the Italian word 'to crush'). When they occur in groups, grace notes can be interpreted to indicate any of several different classes of ornamentation, depending on interpretation. In either case, grace notes occur as notes of short duration before the sounding of the relatively longer-lasting note which immediately follows them. Grace notes, and more general, ornaments can be seen as decorations to, for example example, the main melody.Further reading: `http://en.wikipedia.org/wiki/Grace_note`.

Inverse local tempo is a term used to compare the local speed of a performance to the speed that is prescribed in a score or some other criterium, such as a quantized improvisation. The term is derived from the term local tempo which is defined by the following formula: duration of inter-onset interval in the score divided by the corresponding duration of the inter-onset interval in the score. Hence, the slower the performer plays a certain passage, the higher the inter-onset interval and the lower the local tempo is (since iois in the score are static because score is a static description of the music). Inverse local tempo is, like its name suggests, the inverse of local tempo: performance ioi divided by score ioi. Note that inverse local tempo is not the same as performance ioi, but emphasizes the comparison with an external norm. If the music is performed on basis of a written score, the durations from this score can function as this norm. In the case of improvisation, a quantized version of the improvisation can function as this norm.

Leap is the term used to indicate a relatively big jump in pitch or tone interval. Such a jump is often accompanied by a timing-related phenomenon in music performance. A big leap, for example, can result in a small delay.

Meter, metrical level are the terms used to describe the repetitive pattern of beats in a piece of music. The meter is usually printed in at the beginning of a score with two numbers right above each other. For example 4/4, 3/4 or 6/8 (the upper number denotes how many notes form a bar and the lower number indicates notes of which length should be considered as one counting unit, such as 4 for a fourth note). There are several subdivisions of the bar, which is a unit to describe the length of the pattern one level of weak and strong beat that is repeated. The most salient metrical level, in which one is tempted to tap along with the beat, is called the tactus. Next to this level, there are other levels of repeating weak and strong beats, which always correspond hierarchically with each other. For a more elaborate description, see for example Lehrdahl and Jackendoff (1983).

Note duration is the term used to describe how long a note sounds. Corresponding terms are note onset and offset, indicating the time when a note starts to sound and seizes to sound, respectively. The note duration can be understood in two different time scales corresponding with score (usually rational numbers) and performance (usually real numbers). A note of which the duration in the score is given as 1/8, for example, can be played with a duration of, let's say for the sake of this example, .23 seconds. Note duration is not the same as inter-onset time, inverse local tempo or articulation, although these concepts are closely related. Inverse local tempo emphasizes a relation between performance and score inter-onset times and and inter-onset time is the time-span between two succeeding note onsets. Articulation links duration and inter-onset time in the sense that the sounding parts of notes can , for example, overlap (which the composer can direct in the score with the term legato). In score durations, one often speaks of fourth, eighth, sixteenth notes and so on. Obviously, a fourth note should sound twice as long as an eighths note and four times as long as a sixteenth note when they are in the same score. Many times extra notations such as dots mean something to the score notation. For example, a dotted fourth note should be played one and a half time as long as a fourth note.

Pitch is something perceived by the human ear, as opposed to frequency, the physical measurement of vibration. The note A above middle C played on any instrument is perceived to be of the same pitch as a pure tone of 440 Hz, but does not necessarily contain that frequency or only that frequency. Furthermore, a slight change in frequency need not lead to a perceived change in pitch, but a change in pitch implies a change in frequency. Source and further reading: `http://en.wikipedia.org/wiki/Pitch_(music)`

Phrase is term used to describe parts of the melody which are kind of autonomous entities, like parts of sentences in natural language. A phrase can, just like a sentence from natural language, consist of smaller sub-phrases. Every part as well as the combination of parts can have a timing-related profile in music performance. See section 1.2 for more.

Quantization is the problem of finding the most probable score, given a performance. Quantization finds application in, for example, music programs, where musicians try to render a score by their performances. It can have application in the project described in this thesis, as described in 5.2.

Ragtime See `http://en.wikipedia.org/wiki/Ragtime`

Renaissance See `http://en.wikipedia.org/wiki/Renaissance_music`

Ritard Abbreviation for ritardando which means gradual delay in tempo. Commonly one can expect a musician to play with ritard at the end of a long phrase, comparable with the pause at a comma or dot in a spoken sentence.

Romance See `http://en.wikipedia.org/wiki/Romance_(music)`

Structural unit is the term used in this thesis for an element in the music that sometimes forms an autonomous entity such as a melodic phrase and bar and other times is part of such an autonomous entity like the final part of a phrase. Correspondingly, in this thesis three kinds of structural units are distinguished: phrase-, metrical- (bar), and local units. The concept is used to indicate parts in the music which have a special relation with the timing in the performance. For example, the final part of the 36-phrase in figure 1.5 is marked with a fermata, which indicates that the performer should maintain the duration of the chord in which the phrase ends. At the same time a ritard is expected here. Both phenomena are captured by the structural unit chord-ritard. See also section sec:structure to read more about musical structure. Another obvious structural entity, that can even be derived directly from a written score (using the vertical dashes which mark the its boundaries), is bar. This structural unit is used to capture the timing related to the most important metrical level.

Index

Bibliography

The complete MIDI 1.0 Detailed Specification - Incorporating all Recommended Practices - document version 96.1. The Midi Manufacturers Assosication, 1997.

M. Clynes. Expressive microstructure in music, linked to living qualities. *J. Sundberg (Ed.), Studies of Music Performance*, pages 76–181, 1983.

P. Desain and H. Honing. Does expressive timing in music performance scale proportionally with tempo? *Psychological Research*, 56:285–292, 1994.

P. Desain and H. Honing. The quantization problem: traditional and connectionist approaches. *Understanding music with AI: Perspectives on Music Cognition, Balaban, Ebcioglu, Laske (eds.)*, pages 448–463, 1992.

P. Desain, H. Honing, and H. Heijink. Robust score-performance matching: Taking advantage of structural information. *Proceedings of the 1997 International Computer Music Conference San Francisco: ICMA*, pages 336–340, 1997.

M. Hamanaka, K. Hirata, and S. Tojo. Automatic generation of metrical structure based on gttm. *Proceedings of ICCM 2005*, 2005.

H. Honing. Poco: an environment for analysing, modifying, and generating expression in music. *Proceedings of the 1990 International Computer Music Conference*, pages 364–368, 1990.

F. Lehrdahl and R. Jackendoff. *A generative theory of tonal music.* Cambridge, Mass: MIT Press, 1983.

Norvig. P. *Paradigms of Artificial Intelligence Programming.* Morgan Kaufmann, 1992. ISBN 1558601910.

C. Palmer. Music performance. *Annual Review of Psychology*, 12:115–138, 1997.

D.B. Percival and A.T. Walden. *Spectal Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques.* Cambridge University Press, 1993.

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

W.L. Windor, R. Aarts, Desain P., H. Heijink, and Timmers R. The timing of grace notes in skilled musical performance at different tempi: a preliminary case study. *Psychology of music*, 29:149–169, 2001.

L. Windsor, P. Desain, A. Penel, and M. Borkent. Dissect, a structurally guided method for decomposition of expression in music performance. *Journal of Acoustic Society of America (scheduled for February 2006)*, 2005.