# Experimentally Investigating the Use of Score Features for Computational Models of Expressive Timing

Sebastian Flossmann[1], Maarten Grachten[1], and Gerhard Widmer[1,2]

[1]*Dept. of Computational Perception, Johannes Kepler University, Linz, Austria*
[2]*Austrian Research Institute for Artificial Intelligence, Vienna, Austria*,
{sebastian.flossmann,maarten.grachten,gerhard.widmer}@jku.at

## ABSTRACT

In the expressive performance of music variation of tempo plays a major role in shaping and structuring the piece. We distinguish two aspects of tempo, the current tempo and the timing of individual notes with respect to the current tempo. Those two notions are influenced differently by the characteristics of the performed score. The relation between score and timing/tempo has many facets, one of which we examine more closely in the following. More precisely we provide experimental evidence for the hypothesis that timing is more aptly modeled with score characteristics from a small temporal score context, while tempo modeling profits from a bigger temporal score context.

## I INTRODUCTION

The expressive performance of a piece of music is influenced by a variety of factors depending on the epoch and the score and in large parts on the performing musician and their personal understanding of the piece. Two main factors with which expression as well as musical structure can be made audible are loudness and tempo.

The following experiments investigate the relation between score characteristics and tempo variation in a expressive performance. Tempo itself is considered to consist of two different parts: the *local tempo*, and the *timing*. Local tempo refers to the slowly changing current playing tempo of the performance, used to shape e.g. phrases and final ritardandi [8]. Individual notes, on the other hand, though perceived as played in tempo are almost never played on their exact nominal onsets. Although very small deviations are beyond the means of both human perception and human motor control, the greater deviations are used deliberately to accentuate single notes and create microstructure via anticipations and delays.

As the two notions of tempo differ mainly in temporal scope, it stands to reason that the decision whether to increase or decrease the current tempo, respectively anticipate or delay an individual note are also based on information from different scopes. More precisely the hypothesis that shall be tested in the following is that the timing is related to local characteristics of the score while the tempo relates to more global information.

To test the proposed hypothesis a machine learning algorithm (a Support Vector Machine, for further information see [1]) is trained to predict the tempo and timing curves separately from score characteristics, or *score features*. The features are calculated over small slices of the score symmetrically surrounding the note they refer to. The size of the slices, the *feature scope*, can be varied from note-level to several beats. Hence the influence of the scope on the predictability of the tempo and timing curve and in consequence the suitability of the different scopes for the two curves can be analysed. As a by-product the experiments give insights into which kind of features provide general clues for modeling tempo and timing.

## II RELATED WORK

Some work has been done on modeling performances through score characteristics. As [11] contains an extensive list of references only some of the more prominent ones shall be mentioned here. In [10] Widmer presents a machine learning algorithm that discovers rules governing performance tempo and loudness from score information. The score information used contains rhythmic as well as melodic aspects which are then combined and searched by a combination of sophisticated association rules and clustering methods.

In [5] the authors present a rule system constructed by an "analysis-by-synthesis" approach, a process where a professional musician evaluates tentative rules by judging the produced output. The rules work on a local musical context and describe changes to tempo, loudness and articulation that make up the expressive performance. Most rules describe very local, note-level contexts of indivdual notes, few rules cover entire phrases.

A very recent approach to modeling tempo is [6], where a score feature approach using Hierarchical Hid-

den Markov Models is used to learn expressive tempo and loudness variations.

## III DATA

The experiments are based on a recording of Mozart's piano Sonata KV 279 on a Boesendorfer SE 290 Computer Controlled Grand Piano by the Viennese pianist R. Batik. All played pitches are contained in the data together with their precise on- and offset, loudness and pedaling information. After aligning the performance to a digital representation of the score, all score information like nominal onsets and durations, metrical position, rhythmic context is available for all played notes. W. Goebl provided a four level hierarchical phrase analysis of the piece.

## IV TEMPO AND TIMING

The terms *local tempo* and *timing* refer to two different notions. Local tempo relates to the current speed of the performance measured in beats per minute; timing on the other hand addresses the deviations of individual notes from their expected onset with respect to the local tempo. According to [4] smoothing the instantaneous tempo over 3 beats best fits the human perception of playing tempo. From an analytical point of view this can be identified with the low frequency content of the instantaneous note tempo curve.

The smoothed tempo curve is calculated as follows: first we determine an preliminary tempo value for each soprano note (which we assume to carry the melody) based on inter beat intervals (IBIs). Let $ibi_i$ denote the duration of a 1 beat window placed symmetrically around the onset of the note $s_i$. The *ibi*-tempo is then calculated by $t_i^{ibi} = \frac{60}{ibi_i}$, measured in beats per minute (bpm). By averaging the $t_i^{ibi}$ over a 3 beat window surrounding each soprano note we get smoothed tempo values $t_i^s$ for each soprano note $s_i$.

Subtracting the low frequency content from the original tempo curve, namely $t_i^{ioi}$ calculated directly from the inter onset intervals of the soprano notes, leaves a residual containing only the high frequency changes. The ratio of residual and local tempo represents the timing $t_i^t$ of each note with respect to the current tempo: $t_i^t = \frac{t_i^{ioi} - t_i^s}{t_i^s}$. Figures 1 and 2 show the resulting curves.

This method has one definite drawback. It is a widely accepted fact that phrase boundaries are often marked by a decelerando followed by an accelerando with the phrase boundary somewhere near the slowest point. As a consequence the tempo undergoes a few very fast alterations near this minimum. Although being part of the tempo variation rather than the timing of individual notes those fast

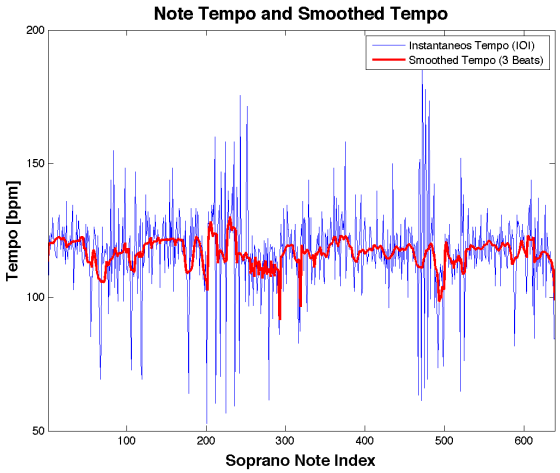


**Figure 1**. The instantaneous tempo and smoothed tempo curves of KV279:1b

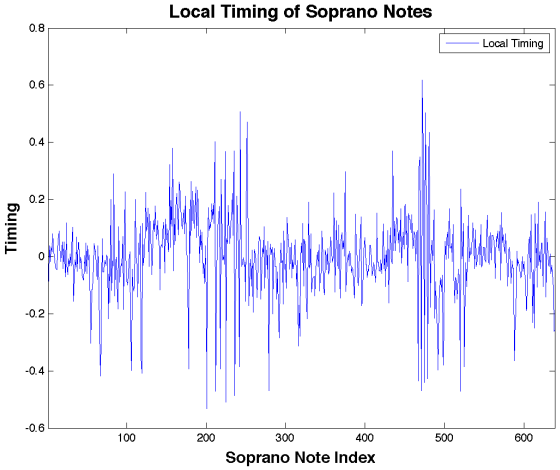


**Figure 2**. Local timing of KV279:1b, relative to the smoothed tempo curve

changes are filtered out in the smoothing process, leaving them in the timing curve. With prior knowledge of the phrases this flaw could be eliminated by weighting the points appropriately but this would limit tempo calculation to problems where phrase information is at hand. Estimating the phrase boundaries beforehand e.g. with a dynamic programming [3] approach could help to solve this problem.

## V SCORE FEATURES

In the following the score features used for the prediction task are described. The same features are used for both tempo and timing prediction. Each of the features can be evaluated at different levels of locality corresponding to

the size of the context window. Let $s_i$ denote the i-th soprano note of the score, $t_i^n$ its nominal onset and $S_i^x$ the set of soprano notes within a window of $x$ beats surrounding $s_i$.

**Melodic Progression (MP)** This features describes the evolution of the melody line by means of an average pitch. To induce temporal dependency the difference in average pitch between two consecutive contexts $S_{i-1}^x$ and $S_i^x$ is is taken as feature value for $s_i$.

**Rythmic Context (Rh)** The rhythmic Context addresses the current liveliness of the piece in the form of a moving average of the nominal note duration over the chosen scope.

**Harmonic Consonance (H)** This feature measures the degree of harmonic consonance of a soprano note with the most probable key for the beat in question. The most probable key is decided by correlating the encountered pitchclasses with key profiles extracted from annotated corpora and assuming the key with the highest correlation. This is a part of Temperley's key finding algorithm described in [9]. The consonance of a soprano note is then measured by the probability of the note given the most probable key using the same key profiles. The value of harmonic consonance for a context is calculated by averaging over all soprano notes within and then taking the difference to the previous context to induce temporal relations.

**Phrase Position (P1-P4)** This features is based on a hierarchical phrase analysis with four levels. Accordingly each soprano note can be assigned four values, describing its position within the current phrase of each phrase level. Following Todd's approach in [7] the positions are calculated relative to the beginning of the phrase. Hence if $s_{i,1}^k$ and $s_{i,n}^k$ are the soprano notes marking the beginning and the end, respectively, of the $i$-th phrase in phrase level $k$ and $s_j$ a soprano note within the phrase the level $k$ phrase position $phr_k$ of $s_j$ is calculated as follows:

$$phr_k(s_i) = \frac{t(s_j) - t(s_{i,1}^k)}{t(s_{i,n}^k) - t(s_{i,1}^k)}.$$

The phrase position is the only feature that can not be made to depend on the scope due to the discrete and not beat related nature of the four hierarchical levels. It is obviuos however that the first phrase level is the most local and the fourth phrase level the most global of the four.

**Metrical Strength (MS)** The accents of the onsets define the metrical structure of the piece, which implies a metrical grid and a segmentation into measures. The metrical grid of a measure consists of several levels of beats, corresponding to different rhythmic values. Every second or third beat of one level is a beat at the immediately higher level. The duple or triple relationships between the levels define different time signatures [9]. The amount of levels at a certain rhythmic value corresponds to the metrical strength of the position. For this feature we associate scope with the resolution of the metrical grid. This seems justified as the taking smallest scope of 0 leads to a grid distinguishing the smallest rhythmic unit occurring in the piece, and so giving the most local information.The grid with the lowest resolution, distinguishing only between beats, is considered to be the widest scope for this feature. Scopes greater than 1 will lead to the same feature values.

**Grace Context (G)** The grace context is a binary feature that indicates if there is a grace note immediately before the chosen context.

**Rest Context (R)** The Rest context specifies the length of a rest immediately preceding the context or is zero if there is no rest.

## VI EXPERIMENT SETUP

As outlined above the main goal is to investigate relations between feature scope and scope of tempo. We look for experimental support for the hypothesis that the timing of individual notes depends on local characteristics while the current tempo relates to more global characteristics of the score.

A Support Vector Machine (SVM) is used for the prediction task, for reasons of high generalization as well as quick training and few structural parameters. A gaussian kernel with a kernel parameter of $0.35$ was found to work best. We used the SVM-KM implementation [2] for Matlab in our experiments. The SVM was trained on the second part of the piano sonata KV279:1 of Mozart (bars 39-100, 639 soprano notes) and tested on the first part (bars 1-38, 393 soprano notes) of the same movement. The quality of prediction is measured by the correlation coefficient between the output and the target curves of the test data.

The experiment is conducted as follows: tempo and timing are both predicted with a scope gradually increasing from $0$ to $5$ beats; for each scope each of the $2^n$ possible feature combinations is tested for the training. In this way the set of features producing the best results as well as the highest prediction quality for the current scope can be discovered. Furthermore although the phrase positions only provide four different levels of globality that can not be related to the chosen scope, by using all available phrase levels in the combination tests the same information can be gathered.

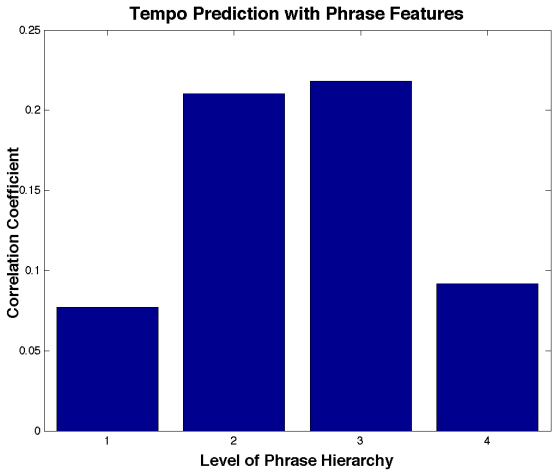


**Figure 3**. Tempo prediction using only single phrasal features

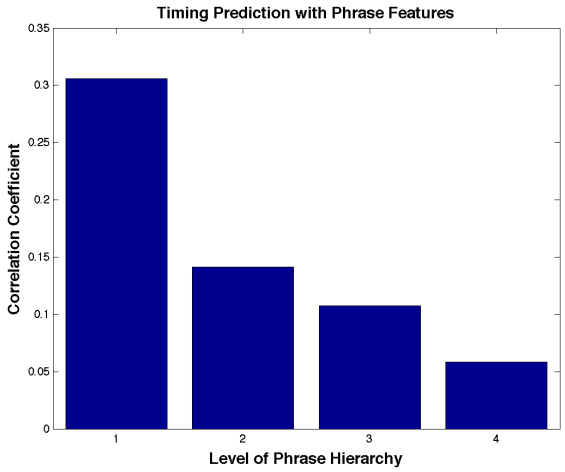


**Figure 4**. Timing prediction using only single phrasal features

## VII RESULTS

### A Single Features

The first issue to be discussed is the predicting quality of single features. The scope, i.e. the size of the score slices measured in terms of beats, is increased from 0 to 5 by steps of 0.5 beats. At each level of scope both curves are predicted using only one feature at a time. Because the phrase level features don't depend on the scope parameter, the results are shown in separate figures (Figures 3 and 4). Figures 6 and 5 shows the remaining non-phrasal features.

On the timing data the most local phrase position, phrase level 1 achieves a correlation coefficient of 0.3076 and clearly outperforms all other phrase levels by more than 100%. On the tempo data phrase levels 2 and 3 perform best with the correlation of phrase level 3 (0.2182) being slightly higher than for phrase level 2 (0.2103).

The picture of tempo and timing prediction with non-phrasal features basically shows the same tendencies. Only two features seem able to model the tempo, the rhythmic context and the melodic progression which peak at a window size of 3 beats (rhythmic context, correlation of 0.5297) respectively 3.5 beats (melodic progression, correlation of 0.2264). The correlations on the timing data are generally lower which is clearly due to the rough nature of the target curve. Very interesting is the peak of the rhythmic context at a context size of 1, which suggests that a small amount of information on the surrounding notes clearly influences the timing of individual notes.

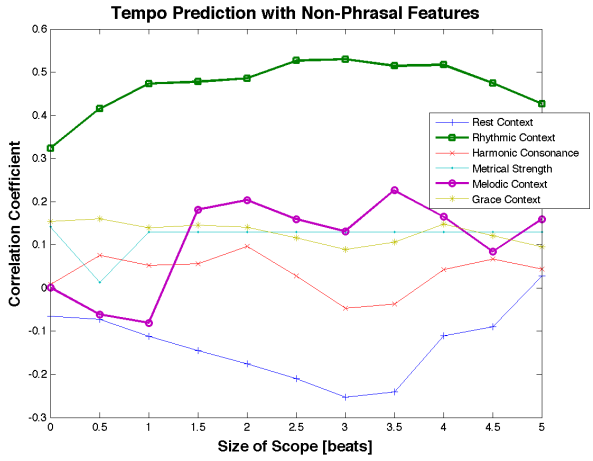


**Figure 5**. Tempo prediction using only single, non-phrasal features

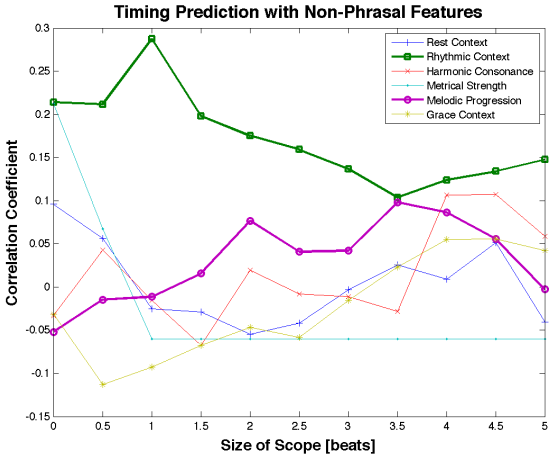


**Figure 6**. Timing prediction using only single, non-phrasal features.

## B Best Feature Combinations

Testing all possible combinations of features also provides the overall best feature set for each scope, which are presented in table B and shall be discussed briefly in the following. The corresponding correlations can be seen in Figure 7. The prediction of timing generally produces lower correlations than the prediction of tempo due to the very fast changing nature of the target curve. As the results suggested a peak between 0.5 and 0.75 this additional scope was also evaluated and produced a very clear peak of 0.53367 for the prediction of the timing. The phrase feature dominating the timing predictions is phrase level 1, describing the smallest possible phrase context. Together with rhythmic context and metrical strength this seems to build a good basis for modeling the timing. The predicted curve using the best cconfiguration of features (Rh,MS,G,P1) at a scope of 0.5 is depicted in 8. As can be seen, although of course the peaks cannot be modeled to their full extent, it models the main tendencies quite well in large parts.

The best feature combinations for tempo prediction are dominated by the rhythmic context and phrase levels 2 and 3. The prediction quality is lowest at a scope of 0 and peaks at a scope of 3.5 beats with a correlation of 0.60637. The predicted smoothed tempo (feature set: Rh, MS,R,MP,P3) is shown in figure 9. Some of the phrases are modeled very nicely, e.g. roughly note 180 - 260 and 280 - 360, while in the beginning the curves seem to behave quite contradictory in small parts. Generally though, the main trends and tendencies are reflected quite well in the predicted tempo.

**Table 1**. Best Combinations of Features

| Scope | Timing | Tempo |
|---|---|---|
| 0 | Rh,R,P1 | Rh,MS,H,P2 |
| 0.5 | Rh,MS,G,P1 | Rh,MS,H,P2 |
| 0.75 | Rh,MS,G,P1 | Rh,MS,P2 |
| 1 | Rh,MS,G,P1 | Rh,MS,P2 |
| 1.5 | Rh,MS,R,H,P1 | Rh,MS,P2 |
| 2 | Rh,MS,R,H,P1 | Rh,MS,P2 |
| 2.5 | Rh,MS,P1 | Rh,H,R,P3 |
| 3 | Rh,MS,P1 | Rh,H,R,P3 |
| 3.5 | Rh,MS,R,G,P1 | Rh,MS,R,MP,P3 |
| 4 | Rh,MS,R,G,P1 | Rh,H,G,R,P2 |
| 4.5 | Rh,MS,G,P1 | Rh,H,G,P3 |
| 5 | Rh,MS,H,P1 | Rh,MP,P3 |

## VIII CONCLUSION

The intention of this research was to investigate the relations between the tempo and timing variations in the performance of a piece of music and the score. The hypothesis was that the local timing can more aptly be de-

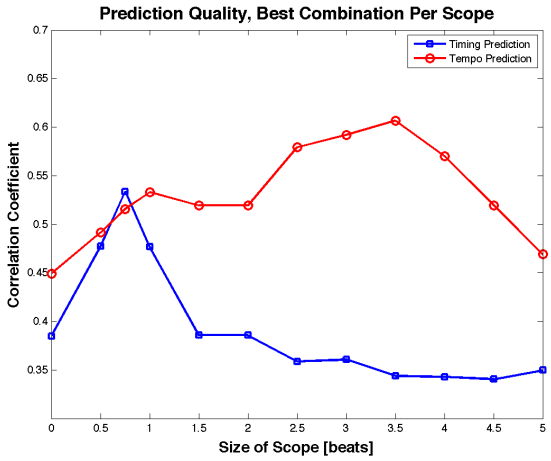


**Figure 7**. Prediction of tempo and timing, best results per scope

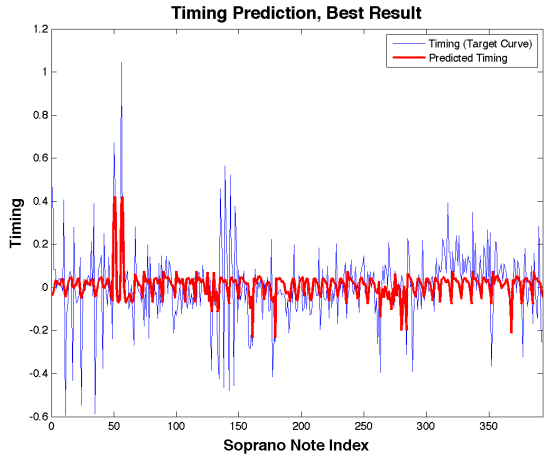


**Figure 8**. Predicted timing with the best found configuration (scope 0.5, Rh,MS,G,P1)

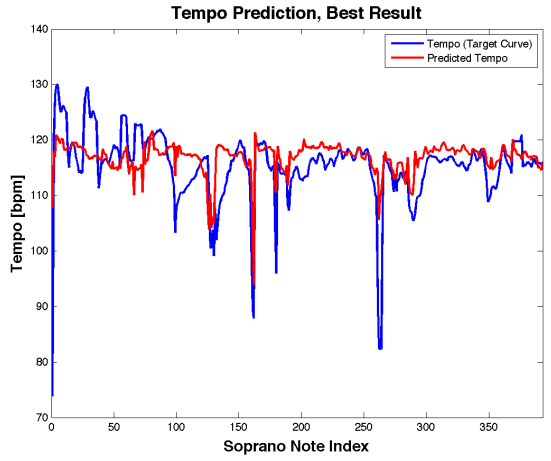


**Figure 9**. Predicted tempo with the best found configuration (scope 3.5, Rh,MS,R,MP,P3)

scribed and modeled by local features while the tempo is more strongly related to large scoped characteristics of the score.

The experiments support this hypothesis. Especially the rhythmic context and the phrase position show a strong relation between the feature scope and the prediction quality. This can be valuable information when working with feature based tempo models. The way we treat the tempo information, splitting it into tempo and timing by low-pass filtering, proved a very suitable way of making this data more easily manageable and interpretable.

## IX ACKNOWLEDGMENTS

This work is funded by the Austrian National Science Fund (FWF) under project number P19349-N15.

## X REFERENCES

[1] BURGES, C. J. C. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery 2*, 2 (1998), 121–167.

[2] CANU, S., GRANDVALET, Y., GUIGUE, V., AND RAKOTOMAMONJY, A. Svm and kernel methods matlab toolbox. Perception Systèmes et Information, INSA de Rouen, Rouen, France, 2005.

[3] CHUAN, C.-H., AND CHEW, E. A dynamic programming approach to the extraction of phrase boundaries from tempo variations in expressive performances. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007), Vienna, Austria* (September 2007), S. Dixon, D. Bainbridge, and R. Typke, Eds., pp. 305–308.

[4] DIXON, S., GOEBL, W., AND CAMBOUROPOULOS, E. Perceptual smoothness of tempo in expressively performed music. *Music Perception 23*, 3 (2006), 195–214.

[5] FRIBERG, A., BRESIN, R., AND SUNDBERG, J. Overview of the kth rule system for musical performance. *Advances in Cognitivie Psychology, Special Issue on Music Performance 2*, 2-3 (2006), 145–161.

[6] GRINDLAY, G., AND HELMBOLD, D. Modeling, analyzing, and synthesizing expressive piano performance with graphical models. *Mach. Learn. 65*, 2-3 (2006), 361–387.

[7] MCANGUS TODD, N. P. The dynamics of dynamics: A model of musical expression. *Acoustical Society of America 91* (June 1992), 3540–3550.

[8] REPP, B. H. Expressive timing in schumann's "träumerei": An analysis of performances by graduate student pianists. *Journal of the Acoustical Society of America 98*, 5 (November 1995), 2413–2427.

[9] TEMPERLEY, D. I. *Music and Probability*. MIT Press, January 2007.

[10] WIDMER, G. Discovering simple rules in complex data: A meta-learning algorithm and some surprising musical discoveris. *Artificial Intelligence 146*, 2 (June 2003), 129–148.

[11] WIDMER, G., AND GOEBL, W. Computational models of expressive music performance: The state of the art. *Journal of New Music Research 33*, No. 3 (2004), 203–216.