Temporal dependencies in the expressive timing of classical piano performances

Maarten Grachten and Carlos Eduardo Cancino Chacón

Abstract

In this chapter, we take a closer look at expressive timing in classical piano performances. We discuss various modeling approaches that attempt to capture how expressive timing is shaped by information present in the written score. Among these, we propose a recurrent basis function model allowing for temporal interactions between intermediate representations of musical score contexts, to learn dependencies of expressive timing on the preceding and following score contexts. We find that this temporal approach predicts expressive timing better than static models using the same basis function representations. Finally, we discuss examples of temporal dependencies that the model has learned.

1 Introduction

As a means of inter-human communication, music is omnipresent, and capable of affecting human states of mind. Although very diverse manifestations of musical behavior can be found throughout humanity, a common modality of musical communication involves composer, performer, and listener roles (Kendall and Carterette, 1990): The composer conceives of a piece of music, and notates the music in the form of a written score, to be interpreted and performed by musicians, directly in front of an audience, or for the purpose of an audio recording,

to be listened by an audience. In this (idealized) modality, the composer encodes his communicative intentions into the music notation. By performing the notated music, the performer recodes the music notation into an acoustic signal, and by listening, the listener decodes this signal into the original intentions of the composer.

What kinds of information are transmitted through music, and how listeners are capable of decoding such information, are fundamental questions that have spawned research for decades. The idea that music is essentially related to motion and physical forces, either literally, or as a metaphor, dates back at least to Truslit (1938). A related account of how the decoding of musical information by the listener may take place is given by Leman (2007). In his view, listeners interpret the form of the acoustic signal by relating it to their action-oriented perception of the world— their *action-oriented ontology*. This ontology emerges from the subject's corporal interaction with the world, and is rooted in motor resonances to physical energy.

The above account explains how listeners interpret what they hear, and as such provides hints as to how music should be performed, when a composer or musician has a specific intention. For example, it entails that if the intention is for the music to express joy, it should be performed in ways that mimic the physical forms that usually accompany joy in the real world, that is vivid, and staccato, rather than slow and legato. The correspondence between the perceived affective character of the music, and certain expressive musical characteristics like tempo, articulation, texture, and dynamics, have been confirmed in numerous studies, in different settings (Juslin and Laukka, 2004; Gabrielsson, 2001).

But musical information is often highly structured, and in addition to the general relations between emotional qualities and expressive characteristics of the music mentioned above, musical expression is strongly related to structural aspects of the music (Clarke, 1988). Notably, musical grouping structure (the hierarchical division of the music into smaller units like *phrases*, and *motifs*) is reflected in tempo and dynamics in the form of arc-like shapes. Todd (1989)

demonstrates that such shapes can be accounted for by a kinematic model, and a similar model has been proposed by Friberg and Sundberg (1999). Another type of musical structure that musicians express through expressive variations is the metrical structure (Sloboda, 1983). These are examples of an important function of musical expression, namely to facilitate the perception of grouping and other types of relevant structure in the music by the listener.

This shows that musical expression is a crucial factor in the interaction between the composer's (or musician's) intentions, and the listener's perception. But musical expression also involves another form of interaction. Istók et al. (2013) argue that proper expressive timing prepares listeners for important musical events, by directing their attention towards those events.

This suggests that expressive parameters, such as variations in timing and dynamics, should not just be regarded as a function of *present* musical events in the score, but also of *future*, and possibly *past* events.

A review of computational models for musical expression of score information (Section 2) reveals that few existing models allow for capturing temporal interactions in the sense described above. We propose a bi-directional variant of recurrent network models as a way of computationally modeling such interactions for expressive timing (Section 3). An experiment on a large corpus of Chopin piano performances shows that this model accounts substantially better for variations in timing than two static models that cannot learn temporal interactions (Section 4). An analysis of the trained models reveals some clear dependencies of expressive timing on past and future events.

2 Data driven methods for learning musical expression

A variety of computational models for various forms of musical expression have been proposed over the past decades. Rather than giving a wide overview, we will

focus on a few data driven models, that are intended to infer regularities in musical expression from data sets of recorded performances. We refer the reader to Kirke and Miranda (2009) for a broader review of computational models of expressive music performance.

Widmer (2002) uses an inductive logic programming approach to discover interpretable performance rules from data. Some of the discovered rules clearly capture some temporal dependencies, such as the shortening/lengthening of notes based on the notated durations of neighboring notes. A drawback of this approach is that information about the musical context of a note must be specified as a set of properties of a note, thus limiting the potential dependencies to be learned to the specified note attributes.

Grindlay and Helmbold (2006) describe a hidden Markov model (HMM) of expressive timing. The model involves a number of hidden states, where each state is associated with a joint probability distribution over the score information and the expressive timing (the observation distribution for that state). During training, the model infers the observation distributions for a set of states, such that sequences of those states account for the observed data as well as possible. Since the observation probabilities only model the dependencies between score information and expressive timing at the same time instance, there is no way to learn the relation between expressive timing and temporally remote score information directly. The transition probabilities between hidden states can in principle capture temporal dependencies between the score features in one state and the expressive timing in another, but since the number of states is limited due to computational constraints, it is unlikely that individual temporal dependencies can be modeled adequately.

Teramura et al. (2008) use Gaussian processes (GPs), a probabilistic kernel method, to render music performances. This approach uses a low level encoding of certain aspects of the musical score, such as the time signature and the dynamics annotations, as well as relative pitch, and an explicit encoding of the melody voice of a piece. Both the encoding of the input data, as well as the GP allow

for modeling temporal dependencies on the input data. This framework is more oriented towards prediction than towards the analysis of the influence of particular score aspects on musical expression. Furthermore, this model does not allow for using future score information.

3 A recurrent basis function model for expressive timing

The recurrent model we describe in this section, is used in combination with *basis function modeling*, a representation scheme for written musical scores, proposed by Grachten and Widmer (2012). Although the basis functions (Section 3.2) do allow for expressing more than strictly local dependencies (for example, the gradual decrease of tempo over the course of a *ritardando*), the shape of these interactions is hard-coded in the basis functions, rather than learned from the data. By combining the basis functions with a recurrent model, it is possible to learn additional, perhaps less obvious interactions between expressive events and score information from the past or future. Before we describe the recurrent model, we give an operational definition of the notion of expressive timing, and we introduce the basis function modeling approach.

3.1 Definition of expressive timing

The notion of expressive timing is ambiguous. Musically speaking, it makes sense to decompose temporal fluctuations in performed music into a *local tempo* component, that represents the rate at which musical units of time (beats) pass, and a *timing*.

In this chapter, we will not make a distinction between tempo and timing. Instead, we will define the parameter of expressive timing based on the inter-onset intervals (IOIs) between consecutive performed notes. The onsets of performed notes that correspond to notes with identical onsets in the score, are averaged.

Deviations of individual note onsets with respect to this average can be modeled as a further expressive parameter, but in this chapter we ignore these deviations.

The actual target to be predicted by the expression model is the ratio of the performed IOI and the corresponding notated IOI. Since we are interested in fluctuations of timing, rather than the overall tempo of the performance, we normalize score and performance IOIs by the total duration of the score (in beats) and performance (in seconds), respectively. Finally, we take the base-2 logarithm of this ratio, such that doubling and halving of tempo corresponds to decrements and increments of the target the same size, respectively.

In the definition given above, expressive timing is represented as a numerical value associated with each IOI, such that the expressive timing of a complete performance takes the form of a series of values over time. Although the expressive timing value actually describes the time interval between two positions in the musical score (two consecutive note onsets), it is usually associated to either the start, or the end of the IOI. In this study, we associate the values to the start of the IOI, such that the expressive timing value at onset i in the score describes the interval between onsets i and $i + 1$ (rather than between $i - 1$ and i).

3.2 Basis function modeling of musical expression

Basis functions are a very general means of representing musical information in a form that allows it to be used to explain and predict expressive parameters, in this case expressive timing. For each onset in the score, they produce a numerical value, which may encode local features, like note duration, metrical position or the pitch of a note, but they may also span longer contexts, for example to denote the range of performance directives such as *piano* (p), *forte* (f), or the position within a slur. Depending on the type of information they encode, basis functions may be mostly zero. For example, the basis function encoding the first beat in a 4/4 time signature evaluates to one only at onsets that fall on the first beat of a 4/4 bar, and to zero everywhere else. When the piece contains no passages in 4/4 time signature, the basis function evaluates to zero over the whole piece (and can

be omitted). Similarly, the basis function encoding *piano* passages is one over the span of the passage (that is, from the position of the *p* to the position of the next, overriding loudness annotation), and zero everywhere else.

3.2.1 Basis functions for embodied music interactions

In the work on basis function modeling of musical expression discussed here, the basis functions are used primarily to encode information extracted from the musical score. Note however that other types of information may just as well be encoded in basis functions. In the context of embodied music cognition, basis function modeling may serve as a methodology to study how physical and physiological factors influence expressive aspects of music performance. For instance, when in addition to recording the acoustic aspects of a music performance, sensors are being used to record corporeal aspects of the performer such as body sway, breathing, heart rate, or skin conductivity, the latter signals can be used (after alignment of the performance to the score) as basis functions to model their relationship to musical expression.

3.2.2 Linear versus non-linear basis models

Given a set of basis functions, the expressive timing in the performance of a piece can be modeled by expressing the timing values as a combination of the basis functions. We call such a function, mapping the values of basis functions to expressive parameter values, a *basis model* of musical expression. In its simplest, *linear* form, a basis model just expresses the timing value at a particular position as a linear combination of the basis functions that are non-zero at that position. In this setup, each basis-function has a positive or negative weight parameter that determines how strongly (and in what direction) that basis-function influences the timing values. The weight parameters can be learned from a set of performances, provided the performances are aligned to their respective musical scores.

The advantages of the linear model are firstly that it is easy to find optimal

weight parameters given a set of performances (by linear least-squares regression), and secondly that it is straight-forward to tell what the model has learned from the data by looking at the weight parameters. For example, a positive weight for the *fermata* basis function (which indicates the presence of fermata in the score) shows that the model has learned that a fermata sign causes an increase in the timing (that is, a lengthening of the note at the fermata).

The linear model also has some important shortcomings. Firstly, the model can only learn linear relationships between the basis-function and the expressive parameter. For instance, in the case of modeling the intensity of notes in a performance, it has been observed that the higher the note, the louder it is played (Grachten and Widmer, 2012). However, this relationship is not strictly linear, but roughly follows an S-curve. In that case a basis function that returns the pitch of a note will help to predict the intensity of the note, but it will tend to overestimate the effect of the *higher–louder/lower–softer* effect for the lowest and highest pitches. A second shortcoming in linear models is that it assumes each of the basis functions to influence the expressive parameter independently of all other basis functions. In reality, it is conceivable that the effect of one particular basis function on expressive timing is different depending on which other basis functions are active at the same time.

These shortcomings can be addressed by using a *non-linear* model, such as a *feed-forward neural network* (FFNN). Using an FFNN in modeling intensity in piano performances, Cancino Chacón and Grachten (2015) demonstrate interactions between *crescendo* and *diminuendo* signs. In particular, they show that the effect of a *crescendo* is reduced in the immediate context of a *diminuendo*.

3.2.3 Static versus temporal basis models

In the FFNN that Cancino Chacón and Grachten (2015) used for modeling expressive dynamics, the relation between the input at time step i (the values of the basis functions) and the output at that time step (the value of the expressive target) is

modeled through an intermediate layer. This *hidden*[1] layer is determined only by the current input, and allows for interactions between current inputs (as Cancino Chacón and Grachten (2015) showed in the case of *crescendo* and *diminuendo* signs), but not for interactions between the states of the model at different positions in the piece. As such, the FFNN model is *static*: there is no notion of the music as a sequence, or a process.

In a *recurrent neural network* (RNN), the state of the hidden layer is not only determined by the input at time step i, but also by the hidden state at time step $i-1$ (which in turn is affected by the hidden state at $i-2$, and so forth). As such, the hidden state of an RNN can be regarded as a representation of the prior context of the input–the musical score–up to the current time step. The preparatory function of musical expression (Istók et al., 2013) suggests that when modeling musical expression, not only the prior score context is relevant, but also the posterior context. In such cases, it can be beneficial to use a *bi-directional* RNN. In this model, one part of the hidden state depends on the prior context, and another part depends on the posterior context, such that the joint parts form a representation of the temporal context of the current time step in both directions.

RNNs can be trained to perform a task, such as predicting expressive timing from the musical score, by presenting a musical score as input to the model (in the form of basis-functions), and comparing the output the model produces to the expressive timing values that were recorded from a real music performance. The weight parameters connecting the input, hidden, and output layers of the model can be optimized for the task by iteratively adjusting the weights proportionally to how strongly their contribution to the predicted output produced a discrepancy with the target output. This method of computing weight updates is called *back-propagation through time* (Rumelhart et al., 1986). For a more elaborate review of RNNs see (Lipton et al., 2015).

[1]The intermediate layer is called *hidden* because its state is not observed as part of the recorded data, but is inferred from that data

4 Experimental model evaluation

In this section, we assess the different types of basis models discussed above, in the context of modeling expressive timing in performances of Chopin piano music. More specifically, we use the linear (LBM) and non-linear (NBM) static models proposed by Grachten and Widmer (2012) and Cancino Chacón and Grachten (2015), respectively, in addition to a third temporal basis model, in the form of a bi-directional RNN taking the same basis function representation as input as the LBM and NBM models. We will refer to this as the *recurrent non-linear basis model* (RNBM).

The purpose of this experiment is to test whether the RNBM model, which can account for expressive timing in terms of *temporal* interactions between the score and expressive timing, brings a benefit over the static models. In addition to a quantitative evaluation of the models, we include a brief review of some qualitative aspects of the trained RNBMs, revealing some of the temporal dependencies the model has learned.

4.1 Data

The Magaloff corpus consists of live performances of the complete piano solo works by Frédéric Chopin, performed by renown pianist Nikita Magaloff during a series of concerts in Vienna, Austria, in 1989 (Flossmann et al., 2010). These concerts were performed on a Bösendorfer SE computer-controlled grand piano, which allows for precise recording of the pieces in a proprietary format by Bösendorfer, and then converted into standard MIDI format. All performances have been aligned to their corresponding musical score. This dataset comprises 155 pieces and over 300,000 performed notes, adding up to almost 10 hours of music.

As stated in Section 3.2, the basis models framework encodes a musical score into a set of numeric descriptors. For this experiment we use a total of 220 basis functions. In many cases, there are multiple functions that encode a single aspect

of the score. Some examples are given in the following list:

- **Dynamics markings**. Bases that encode dynamics markings, i.e., performance directives that describe loudness levels such as *p*, *f*, *crescendo*, etc.
- **Polynomial pitch model**. Proposed in (Grachten and Widmer, 2012), These basis functions represent a third order polynomial model to describe the dependency of dynamics on pitch.
- **Ritardando**. Encoding of markings indicating gradual changes in the tempo of the music such as *ritardando* and *accelerando*.
- **Slur**. Description of $legato$ articulations, which indicate that musical notes are performed smoothly and connected, i.e. without silence between each note.
- **Duration**. Encoding of the duration of a note.
- **Metrical Position**. Representation of the time signature of a piece, and the position of each note in the bar.
- **Fermata**. Encoding of markings that indicate that a note should be prolonged beyond its normal duration.

From the formulation of expressive timing given in Section 3.1, it follows that there is only a single target to predict for each unique onset time in the score (unlike expressive dynamics, where simultaneous notes may have a different intensities). Some of the basis functions (such as those encoding metrical position) are also functions of time. Other basis functions, like those encoding pitch and duration, are defined as functions of notes. For each of the latter type of functions, we average over the values of notes with coinciding onsets, so as to make the basis function representation of the score compatible with the representation of the expressive timing.

4.2 Training and evaluation

4.2.1 Procedure

A 10-fold *cross validation* is used to test the accuracy of the predictions of three different basis models for rendering expressive music performances. This means that each model is trained/tested on 10 different partitions (folds) of the data into training and test sets, such that each piece in the corpus occurs exactly one time in a test set. Each fold consists of on average of 139.5 pieces for training and 15.5 pieces for testing. In order to evaluate the accuracy of the learned models, we report three quality measures. The mean squared error (MSE) measures how close the predictions of the model are to their targets. Secondly, the coefficient of determination R^2, expresses the proportion of variance explained by the model. Lastly, the Pearson correlation coefficient (r) expresses how strongly the model predictions and their respective targets are linearly correlated.

4.2.2 Training of the models

Given a set of training data, all models are trained by minimizing the squared prediction error — the difference between the predicted and the actual expressive timing values. For the LBM, since it is a linear model, this is equivalent to linear regression, and the optimal solution can be found in an exact form. For the non-linear NBM and RNBM, training is not as straight-forward, since these models have several *hyper-parameters* — parameters that are part of the specification of the model, but are not optimized during the training procedure — which are typically determined by heuristics.

For instance, the size of the hidden layers of the NBM and the RNBM must be determined before training. In principle the number of units in the hidden layer should be as small as possible, but not so small as to impede the ability of the model to learn from the training data. We empirically found hidden layers of 20 units to be roughly optimal in this sense.

Furthermore, since the parameters of the NBM and RNBM models are updated

Model	MSE	R^2	r
LBM (Grachten and Widmer, 2012)	0.974	0.026	0.232
NBM (Cancino Chacón and Grachten, 2015)	0.940	0.060	0.249
RNBM	**0.864**	**0.136**	**0.374**

Table 1: Predictive results for IOI, averaged over a 10-fold cross-validation on the Magaloff corpus. A smaller value of MSE is better, while larger r and R^2 means better performance.

iteratively by repeatedly passing over the training data, it is possible that after a certain point *over-fitting* will occur: further training will decrease the error on the training data, but not on newly seen data. To avoid this form of overfitting, we use 28 pieces (around 20%) of the training data for *validation*: Rather than training the model on these pieces, the prediction error of the model for these pieces is used to monitor the training process. As soon as the error on the validation pieces starts to increase, training is terminated.

4.3 Results and Discussion

The average of the three quality measures over the 10 folds for each model type is presented in Table 1. All three measures suggest that the RNBM gives a consistent improvement over both NBM and LBM models. Analysis of variance using linear mixed-effects revealed both a main effect of model on the squared error, and a significant interaction between model and fold on squared error. Because of this interaction, a post-hoc comparison (Tukey's HSD test) was performed per fold, rather than over all folds together. These tests show that the RNBM models are significantly better than both NBM and LBM models on all folds ($p < 0.0001$). In seven folds, NBM performed significantly better than LBM ($p < 0.0001$). In one fold, LBM performed better than NBM ($p < 0.0001$), and in two folds NBM, and LBM were not significantly different.

These results seem to confirm the respective benefits of non-linearity and temporal dependencies for modeling expressive timing in classical piano performances, as described in Sections 3.2.2 and 3.2.3. In the following, we focus on the latter aspect, highlighting learned interactions between the hidden states at different times, a capability that set apart RNBM from LBM and NBM models.

4.3.1 Sensitivity analysis

In order to analyze the effects of past and future score information on the performance of a note, we use a method called *differential sensitivity analysis*. This method quantifies how strongly the output of the model changes in response to changes in each of the basis functions at each time step. Plotting the sensitivity values of a particular basis function at different time steps relative to the current time step yields a kind of *temporal profile* of the basis function, showing how that basis function affects expressive timing in its prior and posterior context. This is done in Figure 1, which shows the model sensitivity with respect to three basis functions, for the models trained in each of the ten folds of the cross-validation. The curves show how an activation of the basis function in the present (time step 0) affects IOIs in the past (negative time steps), and in the future (positive time steps).

The different curves within each of the three plots correspond to the models trained on the 10 different folds of the cross-validation. Although there are differences, the overall shapes of the curves within a plot are very similar, showing that the temporal sensitivity profiles are not due to random factors (such as model initialization) that are irrelevant to the data.

A ranking of all 220 basis functions according to model sensitivity showed that the models are most sensitive to the fermata basis (see Section 4.1). Unsurprisingly, the models have learned from the data that the fermata sign indicates a lengthening of the note. However, the part of the plot before time step 0 shows that the fermata sign also affects the IOIs preceding it, provoking an incremental lengthening of the preceding IOIs up to the fermata sign itself. This result

demonstrates the anticipatory function of expressive timing described by Istók et al. (2013).

A different pattern is exhibited for duration. Here there is a negative sensitivity of performed IOI to the notated duration, implying that long notes are systematically shortened. Interestingly, the surrounding IOIs are lengthened slightly. This suggests that durational contrasts in the score are softened, a phenomenon also reported by Gabrielsson et al. (1983). Although both sharpening and softening of contrasts have been found to be a means of expressing emotions Lindström (1992) in music, the models have not learned sharpening effects from the data. This may be due to the character of the music in the training corpus.

Musical accents have a less marked, and more diffuse effect on IOI . Models from different folds appear to have learned different regularities, although there is a weak tendency to slow down towards an accent, followed by a slight speed up following the accent.

5 Conclusion and future directions

In the musical communication process that takes place between musician and listener, an important function of musical expression (among others) is to facilitate the perception of various forms of musical structure to the listener, for example by guiding anticipatory attention to important events in the music (Istók et al., 2013). This implies that temporal dependencies play an important role in the expression of musical structure.

The RNBM expression model presented in this chapter learns such dependencies from musical data through a hidden layer whose states at different time steps interact. The experiments reported in Section 4 show that the recurrent model predicts expressive timing much better than static versions of the model using the same input representations. This is evidence that temporal dependencies do play an important role in expressive timing. A few concrete instances of learned temporal dependencies have been demonstrated and discussed.

In the context of embodied music cognition, the basis modeling approach employed here may be used to study how corporeal aspects of music performance influence expression, as discussed in Section 3.2.1. An interesting extension of the of RNBM model is to add a dependency of the hidden layer on the previous output of the model. In this way, the hidden state of the model does not only represent the context of the musical score, but also the performance (such that the same musical situation may be represented differently, depending on the way it is performed). Such a model may offer a more complete view on the various types of interactions between score and performance context that shape musical expression.

References

Cancino Chacón, C. E. and Grachten, M. (2015). An evaluation of score descriptors combined with non-linear models of expressive dynamics in music. In Japkowicz, N. and Matwin, S., editors, *Proceedings of the 18th International Conference on Discovery Science (DS 2015)*, Lecture Notes in Artificial Intelligence, Banff, Canada. Springer.

Clarke, E. F. (1988). Generative principles in music. In Sloboda, J., editor, *Generative Processes in Music: The Psychology of Performance, Improvisation, and Composition*. Oxford University Press.

Flossmann, S., Goebl, W., Grachten, M., Niedermayer, B., and Widmer, G. (2010). The Magaloff Project: An Interim Report. *Journal of New Music Research*, 39(4):363–377.

Friberg, A. and Sundberg, J. (1999). Does music performance allude to locomotion? a model of final ritardandi derived from measurements of stopping runners. *Journal of the Acoustical Society of America*, 105(3):1469–1484.

Gabrielsson, A. (2001). Emotions in strong experiences with music. In Juslin,

P. N. and Sloboda, J. A., editors, *Music and emotion: Theory and research*, pages 431–449. Oxford University Press.

Gabrielsson, A., Bengtsson, I., and Gabrielsson, B. (1983). Performance of musical rhythm in 3/4 and 6/8 meter. *Scandinavian Journal of Psychology*, 24(1):193–213.

Grachten, M. and Widmer, G. (2012). Linear basis models for prediction and analysis of musical expression. *Journal of New Music Research*, 41(4):311–322.

Grindlay, G. and Helmbold, D. (2006). Modeling, analyzing, and synthesizing expressive piano performance with graphical models. *Machine Learning*, 65(2–3):361–387.

Istók, E., Friberg, A., Huotilainen, M., and Tervaniemi, M. (2013). Expressive timing facilitates the neural processing of phrase boundaries in music: Evidence from event-related potentials. *PLoS ONE*, 8(1).

Juslin, P. N. and Laukka, P. (2004). Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research*, 33(3):217–238.

Kendall, R. and Carterette, E. (1990). The communication of musical expression. *Music Perception*, 8(2).

Kirke, A. and Miranda, E. R. (2009). A survey of computer systems for expressive music performance. *ACM Computing Surveys (CSUR)*, 42(1):3.

Leman, M. (2007). *Embodied Music Cognition and Mediation Technology*. MIT Press.

Lindström, E. (1992). 5 x "oh, my darling clementine". the influence of expressive intention on music performance. Department of Psychology, Uppsala University.

Lipton, Z. C., Berkowitz, J., and Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning. *CoRR*, abs/1506.00019.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(9):533–536.

Sloboda, J. A. (1983). The communication of musical metre in piano performance. *Quarterly Journal of Experimental Psychology*, 35A:377–396.

Teramura, K., Okuma, H., Taniguchi, Y., Makimoto, S., and S., M. (2008). Gaussian process regression for rendering music performance. In *Proceedings of the 10th International Conference on Music Perception and Cognition (ICMPC)*, Sapporo, Japan.

Todd, N. (1989). A computational model of rubato. *Contemporary Music Review*, 3 (1).

Truslit, A. (1938). *Gestaltung und Bewegung in der Musik*. Chr. Friedrich Vieweg, Berlin-Lichterfelde.

Widmer, G. (2002). Machine discoveries: A few simple, robust local expression principles. *Journal of New Music Research*, 31(1):37–50.

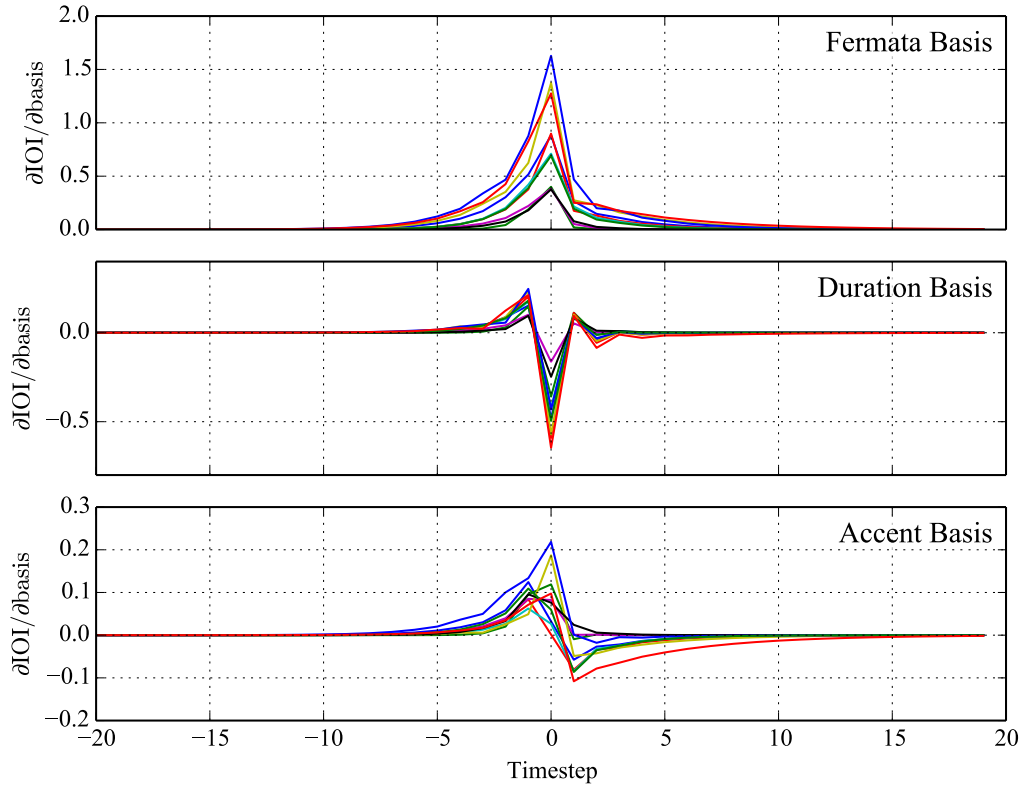

Figure 1: Sensitivity of IOI predictions to Fermata, Duration and Accent basis functions, for the ten models obtained from ten-fold cross-validation; Positive values on the vertical axis denote a lengthening of IOI (tempo decrease), negative values a shortening (tempo increase); The curves show the effect of the activation of the basis function at $i = 0$ on IOI values in the past (negative time steps), and future (positive time steps)