Analysis and prediction of expressive dynamics using Bayesian linear models

Maarten Grachten, Carlos Eduardo Cancino Chacón, and Gerhard Widmer

Austrian Research Institute for Artificial Intelligence
http://www.ofai.at/~maarten.grachten

Abstract. We present the first version of a probabilistic linear basis model for expressive dynamics in music. The model extends prior work by Grachten and Widmer [7]. Contrary to the original approach, this model allows for both specifying musical knowledge, and for modeling multiple distinct performances of the same piece. We show that in its current, minimalist form, the new approach performs on a par with the original model in terms of predictive accuracy. Furthermore, a novel set of basis-functions is evaluated, to model dynamics annotations such as *(de)crescendo* in a context-aware manner.

Keywords: Musical expression, probabilistic linear basis modeling

1 Introduction and related work

Expressive interpretation of notated classical piano music is a complex human skill, that involves extensive practice, and substantial tacit knowledge. It is generally agreed that expressive interpretations serves two main communicative functions. Firstly, musicians tend to shape their performance to communicate musical structure to the listener. The second important function of musical expression is to communicate affect [9,4]. Musicians can play music in a way that amplifies a particular mood that may be inherent in the musical content, but they may also choose to impose their own emotional intentions in the music they perform. Extensive overviews of research on the production, perception, and modeling of musical expression are given in [6].

Computational models have been proposed for a variety of factors that shape musical expression. These models may serve mainly analytical purposes [13,14], mainly predictive purposes [12], or both [8,4,7]. Other models attempt to provide intuitive control over musical expression [3,2]. Of the works mentioned, both [12] and [8] are of special relevance to the work presented here, since they also present probabilistic approaches to modeling musical expression. Teramura et al. [12] use Gaussian processes (GP), a probabilistic kernel method, to render music performances. The *equivalent kernel* in GP uses gaussian basis functions of the input data, instead of basis functions that model specific characteristics of dynamic performances. For this reason, it seems better suited for prediction, than for analysis of the influence of particular score aspects on musical expression. Grindlay and Helmbold's ESP model [8] consists in a hierarchical hidden

Markov model (HHMM) to model the distribution between score features and expressive parameters (tempo). An interesting contribution in their work is the use of an "entropic prior", that favors low entropy distributions over the model parameters, which supposedly results in more interpretable models. Furthermore, they compare model parameters trained on performances of different performers to characterize differences between performers.

Most computational models focus on implicit factors that influence expression like those discussed above. However, to aid interpretation, composers often annotate their music with explicit directives for the dynamics, tempo, and articulation of the performance. Subsequently, editors may also add such annotations, mostly for didactic purposes, rather than to express artistic intentions. To date, few modeling approaches attempt to give an account of musical expression that incorporates directives explicitly written in the score. Repp considers *crescendo* and *decrescendo* signs in a study of dynamics in the opening of a Chopin Etude [11], but this constitutes a case study rather than a generic model. Grachten and Widmer [7] propose a computational model for musical expression that models dynamics annotations (such as *(de)crescendo*, *piano*, *forte*, *sforzato*) explicitly as basis functions, which are combined linearly to model expressive dynamics. Their approach using least squares (LS) regression, although not explicitly formulated in a probabilistic way, is equivalent to a maximum likelihood estimation of the parameters, assuming the note intensity values are normally distributed, given the musical score and the model parameters [1].

In this paper, we propose a probabilistic formulation of the model by Grachten and Widmer, where the parameters are estimated using a maximum a posteriori approach. This formulation makes the model much more flexible. On the one hand, it is possible to incorporate prior knowledge in the form of prior distributions over the model parameters. On the other hand, this formulation alleviates the restrictive assumption of the targets being normally distributed. This is an important step in order to model the fact that there may be multiple distinct ways to perform music. In addition to the formulation of the Bayesian linear basis model for music expression, we experiment with a new type of basis functions to represent *crescendo/decrescendo* annotations, that take in to account the context in which the annotation occurs.

The work presented should be regarded as a proof-of-concept of the probabilistic extension of Grachten and Widmer's approach, rather than a final model for musical expression. We have not yet taken full advantage of the new model by specifying a prior information (e.g. encoding that note intensities tend to increase during a *crescendo* and decrease during a *decrescendo*), or by using a non-normal density function for modeling the distribution probabilities of the note intensities. Nevertheless, we show that the new approach in its simplest form performs on a par with the original model.

Section 2 provides an explanation of the use of basis functions to represent dynamics annotations, as well as pitch information (as in [7]). In Section 3, we formulate the model, and describe how the parameters can be computed given training data. Using a data set of performances of Chopin's piano music by

Category	Examples	Basis function
Constant	*f, p, dolce*	step
Gradual	*crescendo, diminuendo, calando*	ramp + step
Impulsive	*fz, sfz, fp*	impulse

Table 1. Categories of dynamics markings

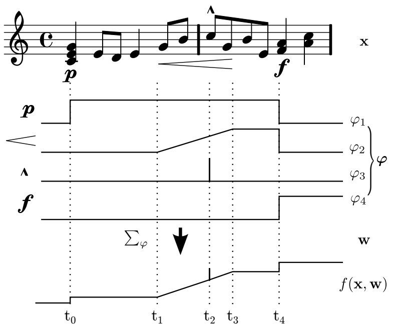

Fig. 1. Schematic view of note intensities as a weighted sum $f(\mathbf{x}, \mathbf{w})$ of basis functions $\boldsymbol{\varphi}$, representing dynamic annotations

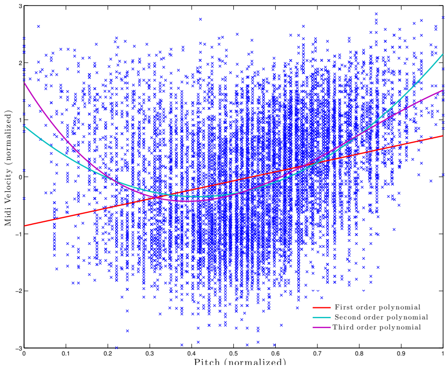

Fig. 2. Dependency of dynamics and pitch in Magaloff's performances of Chopin piano pieces reported in Section 4.1

Nikita Magaloff, we compare the predictive accuracy of the model to that of [7], in Section 4. The results are discussed in Section 5, and conclusions and future work follow in Section 6.

2 Basis Functions

The use of basis functions to represent dynamics annotations in musical scores follows naturally from the fact that these annotations typically have a time range over which they take effect, rather than a single position in time. For example, a *piano* annotation specifies a constant, relatively low loudness level, that is in effect until another such directive occurs, for example a *forte* (implying a relatively high loudness level). Gradual increases and decreases of loudness are indicated by *crescendo* and *decrescendo*, respectively. A third class of annotations concerns only a single position in time (or even a single note), typically to describe an sudden accent (e.g. *sforzato*). Each of these three categories of dynamics annotations has each own basis function, which is intended to model the effect of an annotation in a schematic way (see table 2). Figure 2 illustrates the idea modeling loudness as a weighted sum of basis functions schematically.

The notion of basis functions is very general, and can be designed to capture any aspect of the score. In particular, by defining basis functions that map score notes to various powers of their midi pitch, it is possible to capture the effect of pitch on note intensity in the form of a *polynomial pitch model* [7]. Figure 2 shows a scatter plot of note intensity versus MIDI pitch, for the data set used in Section 4, together with polynomials of different degrees to model the relationship.

In this paper, we define a new set of basis functions, intended to differentiate between gradual loudness annotations in different loudness contexts. We do this by combining each gradual annotation with its preceding and succeeding loudness level, for example *p* → *crescendo* → *mf*, or *f* → *diminuendo* → *mf*.

3 Bayesian linear regression

As already discussed in the introduction, in previous work [7], the estimation of the weights $\mathbf{w}$ that mix the basis models was performed using LS regression. In this paper we use a Bayesian approach, namely, maximum a posteriori (MAP) estimation to compute the weights $\mathbf{w}$. This approach can be formally described as follows:

Given a musical score, represented as a list of N notes $\mathbf{x} = (x_1, \ldots, x_N)^T$ and a set of K predefined basis functions $\boldsymbol{\varphi} = (\phi_1, \ldots, \phi_K)^T$, a sequence of N target values $\mathbf{y} = (y_1, \ldots, y_N)^T$ (e.g. loudness) can be modeled as a linear combination of the basis functions plus noise $\boldsymbol{\epsilon}$ as

$$\mathbf{y} = \boldsymbol{\Phi}\mathbf{w} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\Phi}$ is a $N \times K$ matrix with elements $\Phi_{ik} = \phi_k(x_i)$ and $\mathbf{w}$ is a vector of K weights. If we consider $\boldsymbol{\epsilon}$ as a zero mean Gaussian random variable with covariance $\boldsymbol{\Sigma} = \beta^{-1}\mathbf{I}$, and assume that every sample $y \in \mathbf{y}$ is independent and identically distributed, the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$ is

$$p(\mathbf{y} \mid \mathbf{x}, \mathbf{w}) = \prod_{n=1}^{N} \mathcal{N}(y_n \mid \mathbf{w}^T \boldsymbol{\varphi}(x_n), \beta^{-1}). \tag{2}$$

Since we are not seeking to model the distribution of the input variables $\mathbf{x}$, and thus, $\mathbf{x}$ will always appear as a set of conditioning variables, in the rest of this paper $\mathbf{x}$ will be dropped from the conditional distributions, to avoid cluttered notation.

Using a Bayesian interpretation, we assume that the weights itself have a prior distribution $p(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid \mathbf{m}_0, \mathbf{S}_0)$, where $\mathbf{m}_0$ and $\mathbf{S}_0$ are the mean and covariance respectively. Using this prior distribution and the conditional distribution of $\mathbf{y}$ given the weights $\mathbf{w}$ from Eq. (2), it follows from Bayes' theorem that that the posterior probability of the weights $\mathbf{w}$ given the targets $\mathbf{y}$ is also a Gaussian distribution, i.e.

$$p(\mathbf{w} \mid \mathbf{y}) = \mathcal{N}(\mathbf{w} \mid \mathbf{m}_N, \mathbf{S}_N), \tag{3}$$

where the mean and covariance are given respectively by

$$\mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}^T\mathbf{y}) \qquad \text{and} \qquad \mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}. \tag{4}$$

As a simplification, we assume $\mathbf{m}_0 = \mathbf{0}$ and $\mathbf{S}_0 = \alpha^{-1}\mathbf{I}$. In this case, the posterior log-likelihood takes the form

$$\log p(\mathbf{w} \mid \mathbf{y}) = -\frac{\beta}{2}\sum_{n=1}^{N}(y_n - \mathbf{w}^T\boldsymbol{\varphi}(x_n))^2 - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} + \text{const.} \tag{5}$$

In order to maximize the posterior log-likelihood, we take the gradient with respect to the weights, i.e.

$$\nabla \log p(\mathbf{w} \mid \mathbf{y}) = \beta\sum_{n=1}^{N}(y_n - \mathbf{w}^T\boldsymbol{\varphi}(x_n))\boldsymbol{\varphi}(x_n)^T - \alpha\mathbf{w}. \tag{6}$$

In this way, the weights that maximize the posterior log-likelihood are calculated by solving $\nabla p(\mathbf{w} \mid \mathbf{y}) = \mathbf{0}$, which results in

$$\mathbf{w}_B = \left(\frac{\alpha}{\beta}\mathbf{I} + \boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^T\mathbf{y}. \tag{7}$$

The hyper-parameters α and β can be computed using the *evidence approximation* algorithm (also known as *type 2 maximum likelihood*) [1].

4 Experiments

In order to demonstrate how the model is able to account for aspects of expressive dynamics, a two part experiment was conducted. Using a set of precisely measured performances by a single professional pianist, we first evaluate the ability of the model to explain the expressive dynamic variations using various combinations of basis functions. Then, we test how well the model generalizes to unseen data.

The following abbreviations are used to refer to the different sets of basis functions: PIT is the polynomial pitch model described in [7]; DYN are the dynamics annotations without context, as in [7]. DYN$_c$ represent the dynamics annotations with context (see Section 2).

4.1 Data Set

We use the Magaloff corpus [5], which consists of live performances of the complete Chopin piano works as played by the Russian-Georgian pianist Nikita Magaloff (1912-1992). These performances were recorded in a series of concerts in Vienna, Austria in 1989, using a Bösendorfer SE computer-controlled grand piano [10]. The data was converted into standard MIDI format, where note intensities are represented by MIDI velocity.

Magaloff was known for using manuscripts as scores, but we are uncertain as to the exact version. In this work, the dynamics markings are obtained by optical music recognition (OMR) from the scanned musical scores from the Henle Urtext Edition. Although the OMR program used (SharpEye) does transcribe

	LS				Bayesian			
	r		R^2		r		R^2	
Basis	ave.	std.	ave.	std	ave.	std	ave.	std
DYN	0.516	0.110	0.275	0.107	0.516	0.110	0.276	0.107
DYN_c	0.576	0.095	0.339	0.107	0.576	0.095	0.339	0.107
PIT	0.416	0.033	0.174	0.028	0.416	0.034	0.174	0.028
DYN+PIT	0.622	0.096	0.393	0.112	0.621	0.096	0.393	0.113
$\mathrm{DYN}_c + \mathrm{PIT}$	0.675	0.092	0.462	0.118	0.675	0.092	0.461	0.118

Table 2. Goodness-of-fit of the model over performances of four Chopin piano pieces. See section 4 for abbreviations

dynamics annotations, the transcriptions are not always reliable (false negatives, incorrect positioning). At the time of writing, the dynamics annotations have been manually corrected for four pieces: Op. 15 No. 1, Op. 27 No. 2 (Nocturnes), Op. 28 No.17 (Prelude), and Op. 52 (Ballade).

4.2 Goodness-of-fit of the dynamic representation

For the first part of the experiment, we use the four corrected pieces from the Magaloff corpus, to evaluate the ability of the model to explain expressive dynamic variation in musical performances. As a quantifier of the goodness of fit, we use r, the Pearson correlation coefficient and R^2, the coefficient of determination. The correlation coefficient denotes how strongly the observed dynamics and the dynamics proposed by the model correlate, while R^2 expresses the proportion of variance explained by the model. Table 2 shows a comparison of the observed expressive dynamics and the dynamics proposed by the different models and sets of basis functions. We show the average and standard deviation of both r and R^2 for the LS approximation, as well as for the new Bayesian approach.

4.3 Predictive accuracy

To evaluate the accuracy of the predictions of the trained model, a leave-one-out cross validation over a total of 151 pieces was performed. The model was trained with 150 pieces, and then it was used to predict the dynamics of the remaining piece. Table 3 shows the accuracy of the model in this scenario using again as a quality measures the Pearson correlation coefficient r and the coefficient of determination R^2 for both the LS and the Bayesian regressions and the different sets of basis functions.

5 Discussion

The results for both goodness-of-fit and predictive accuracy show that the Bayesian approach performs on a par with the LS regression. This is an expected result,

	LS				Bayesian			
	r		R^2		r		R^2	
Basis	ave.	std.	ave.	std	ave.	std	ave.	std
DYN	0.181	0.204	0.073	0.086	0.181	0.204	0.073	0.086
DYN_c	0.185	0.198	0.072	0.083	0.185	0.198	0.072	0.083
PIT	0.381	0.145	0.166	0.096	0.381	0.145	0.166	0.096
DYN+PIT	0.431	0.145	0.207	0.113	0.431	0.145	0.207	0.113
$\text{DYN}_c + \text{PIT}$	0.420	0.149	0.198	0.113	0.419	0.150	0.198	0.113

Table 3. Predictive accuracy in a leave-one-out scenario over performances of 151 Chopin piano pieces. See section 4 for abbreviations

since we assumed a zero mean Gaussian distribution for the prior probabilities over the weights. During the realization of the experiments, we noted that the hyper parameter α (the precision of $p(\mathbf{w})$) tends to be very small, suggesting that the prior probability is non-informative, and therefore assuming that the priors have a centered unimodal distribution could be an oversimplification.

We can also see that the use of a more sophisticated basis functions for modeling the dynamics DYN_c do not increase the predictive accuracy. This suggests that there may not be enough training data to significantly represent all possible basis functions (ca. 2500). In case of goodness of fit, the over-modeling provided by the basis functions DIN_c, presents an increase of ca. 12% in the correlation coefficient, and almost 22% more explained variance when compared to the more general basis DYN. Nevertheless this does not reflect in the joint basis $\text{DYN}_c + \text{PIT}$ and $\text{DYN} + \text{PIT}$, where the results are almost identical.

6 Conclusion and future work

In this paper, a fully probabilistic Bayesian approach for analyzing and predicting musical expression using a linear basis model was presented and evaluated. The results show that in its current form, this approach performs almost identically to an LS regression, since under the current assumption of a zero-mean gaussian prior distribution of the weights, the prior is not informative. Nevertheless, the probabilistic formulation has some strong advantages over a least squares approach. In particular, it can take advantage of of musical knowledge as prior information, and it can account for multiple distinct performances of a musical piece, for instance by using mixtures of gaussians rather than a single gaussian to represent the distribution of model parameters.

A newly proposed set of basis functions that model gradual dynamics annotations in a context-aware fashion, although musically justifiable, failed to improve predictive accuracy. An explanation for this is that the context-aware representation of the annotations creates a higher-dimensional, and thus less densely populated data space, in which it is harder to generalize to unseen data.

An obvious next step is to model both the noise and the prior probabilities of the weights by other than Gaussian distributions, for example Gaussian Mixture

Models. Combining this approach with a predictive distribution, it would be possible to render distinct musically acceptable performances of the same piece, rather than a single most likely performance.

Acknowledgments

This work is supported by European Union Seventh Framework Programme, through the Lrn2Cre8 (FET grant agreement no. 610859) and the PHENICX (grant agreement no. 601166) projects.

References

1. C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer Verlag, Microsoft Research Ltd., 2009.
2. R. Bresin and A. Friberg. Emotional coloring of computer-controlled music performances. *Computer Music Journal*, 24(4):44–63, 2000.
3. S. Canazza, G. De Poli, C. Drioli, A. Rodá, and A. Vidolin. Modeling and control of expressiveness in music performance. *Proceedings of the IEEE*, 92(4):686–701, april 2004.
4. G. De Poli, Canazza S., A Rodà, A. Vidolin, and P. Zanon. Analysis and modeling of expressive intentions in music performance. In *Proceedings of the International Workshop on Human Supervision and Control in Engineering and Music*, Kassel, Germany, September 21–24 2001.
5. S. Flossmann, W. Goebl, M. Grachten, B. Niedermayer, and G. Widmer. The Magaloff Project: An Interim Report. *Journal of New Music Research*, 39(4):363–377, 2010.
6. A. Gabrielsson. Music performance research at the millennium. *The Psychology of Music*, 31(3):221–272, 2003.
7. M. Grachten and G. Widmer. Linear basis models for prediction and analysis of musical expression. *Journal of New Music Research*, 41(4):311–322, 2012.
8. G. Grindlay and D. Helmbold. Modeling, analyzing, and synthesizing expressive piano performance with graphical models. *Machine Learning*, 65(2–3):361–387, 2006.
9. P. N. Juslin and P. Laukka. 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, 2004.
10. R. A. Moog and T. L. Rhea. Evolution of the Keyboard Interface: The Bösendorfer 290 SE Recording Piano and the Moog Multiply-Touch-Sensitive Keyboards. *Computer Music Journal*, 14(2):52–60, 1990.
11. B. H. Repp. A microcosm of musical expression: II. Quantitative analysis of pianists' dynamics in the initial measures of Chopin's Etude in E major. *Journal of the Acoustical Society of America*, 105(3):1972–1988, 1999.
12. K. Teramura and H. Okuma. Gaussian process regression for rendering music performance. In *Proceedings of the 10th International Conference on Music Perception and Cognition (ICMPC)*, Sapporo, Japan, 2008.
13. G. Widmer. Machine discoveries: A few simple, robust local expression principles. *Journal of New Music Research*, 31(1):37–50, 2002.
14. W. L. Windsor and E. F. Clarke. Expressive timing and dynamics in real and artificial musical performances: using an algorithm as an analytical tool. *Music Perception*, 15(2):127–152, 1997.