Modelling Gestures in Music Performance with Statistical Latent State Models

Taehun Kim and Stefan Weinzierl
Audio Communication Group, TU Berlin
Einsteinufer 17c, 10587 Berlin
Germany
t.kim@mailbox.tu-berlin.de, stefan.weinzierl@tu-berlin.de

ABSTRACT

We discuss try to identify "gestures" in music performances by observing patterns in both compositional and expressive properties, and by modelling them with a statistical approach. Assuming a finite number of latent states on each property value, we can describe those gestures with statistical latent state models, and train them by unsupervised learning algorithms. Results for several recorded performances indicate that the trained models could identify the gestures observed, and detect their boundaries. An entropy-based measure was used to estimate the relevance of each property for the identified gestures. Results for a larger corpus of recorded and annotated musical performances are promising and reveal potential for further improvements.

Keywords

Musical gestures, performance analysis, unsupervised machine learning

1. INTRODUCTION

Musical ideas and concepts are realised with sound properties varying over time. Some of the important properties are pitch, rhythm, timbre, tempo and loudness, and we often observe particular patterns in those sequences. Such a pattern serves as a medium conveying specific messages and emotions, so it can be understood as a sonic "gesture" [3]. A music performance therefore can be understood as a sequence of "gestures". Hence, both for the analysis and for the re-synthesis of musical performances, the modelling of musical gestures seems to be a crucial element.

A gesture in a music performance is implemented with compositional and expressive properties, in which certain patterns are observed. Figure 1 illustrates examples of such pattern repetitions and variations. The simplest one is an exact repetition such as the rhythmic pattern in Figure 1a. In this case, the repeated pattern can be perceived as a "gesture", and we can regard the boundaries of those repetitions as the gesture boundaries. A gesture can, however, still be recognized if patterns are not exactly repeated, but with certain variations, such as in Figure 1b, where only the half and quarter note at the end of each phrase constitute the boundary of each gesture.

Different properties are observable in a music performance,

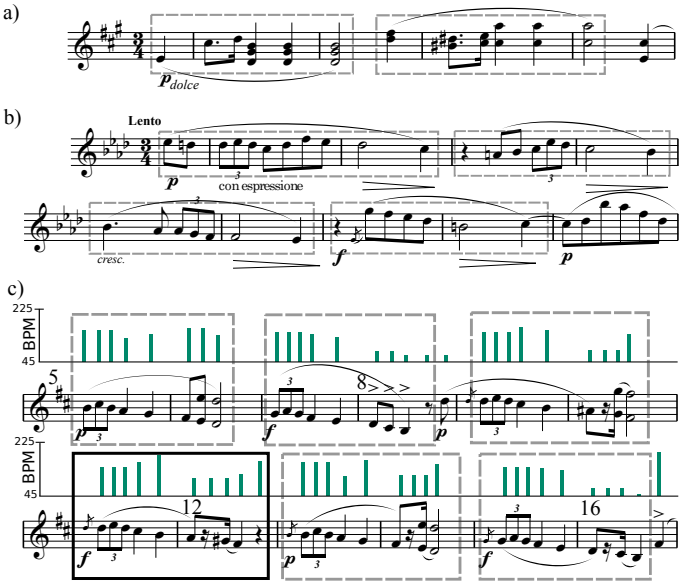

Figure 1: a) F. Chopin, Prelude Op. 28 No. 7, the first 4 bars. b) F. Chopin, Waltz Op. 69-1, the first 9 bars. c) F. Chopin, Mazurka Op. 30-2, bars 5 - 16, performed by H. Czerny-Stefanska found in the CrestMuse PEDB (ID: cho-mzk019-czern-y). Boxes with dotted-line indicate "gestures" observed in the performances.

and those properties can have different effects on the identification of a gesture. In Figure 1a it is easy to find particular patterns in rhythm, but difficult to find ones in pitch. In this piece therefore the rhythm property seems to be more relevant for the gesture identification. Patterns observed in compositional properties are closely related to musical phrases, but patterns in of expressive properties are not always match those phrases. In bars 11-12 of Figure 1c we have observed a sudden slow-down of tempo in the middle of the phrase, and such an ambiguity can easily be found in many other performances.

Rink et al. demonstrated how to identify tempo and loudness gestures by a Self Organizing Map (SOM) based clustering of the expressive properties in every three beats [6]. They identified gestures in 29 different performances of a Chopin *mazurka* performed by different pianists, and analysed them with those identified gestures. In consideration of the characteristics of a mazurka piece, analysing gestures on every three beats seems to be a reasonable assumption, but the length of a gesture is generally variable, so it requires a gesture boundary detection method for a more general analysis task. There are several works focused on

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
NIME'13, May 27 – 30, 2013, KAIST, Daejeon, Korea.
Copyright remains with the author(s).

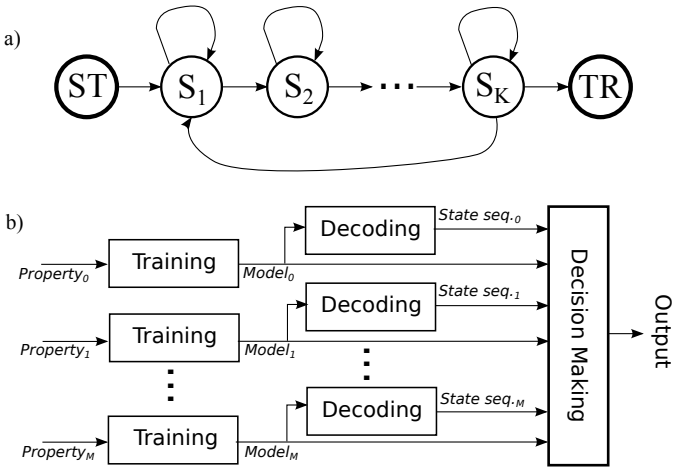

Figure 2: a) State transition diagram. ST and TR represent the beginning and the end of the performance, respectively. b) Overview of a multi-stream approach. The effectiveness measure is used for the decision making.

phrase boundary detection. The Preference Rules [7] and the Local Boundary Detection Method (LBDM) [1] utilize rules defined with compositional properties. Local Maximum Detection [2] attempts to find phrase boundaries by detecting local maxima in expressive property sequences. IDyOM [4] proposes a entropy-based phrase boundary detection method, and shows that a statistical approach would be useful for the detection task. Those works discuss, however, the detection of specific phrases found in musical compositions, but not of underlying models of musical "gestures" as a more abstract representation.

In this paper, we present an approach how to identify gestures in music performance with statistical latent state models. We discuss how to capture pattern repetitions and variations found in all property sequences with a Hidden Markov Model (HMM) based unsupervised learning, and how to estimate the relevance of each property for the identification of gestures. For the evaluation of the model, we analyse the statistical models trained from selected performances, and show the estimated gesture boundaries in those performances. In addition, we show how different interpretations can be analysed with our model, and report test results for a larger corpus.

2. MODELLING APPROACH

2.1 Gesture Hidden Markov Model

A gesture can be described with pitch, duration, tempo and loudness varying over time. We assume that there is a finite number of latent states that represent the beginning, middle and end of a gesture, and that all property changes are observable under specific latent states. A latent state on the n-th note X_n could have K different states such as $X_n \in \{\mathrm{S}_1, \mathrm{S}_2, \ldots, \mathrm{S}_K\}$. A property observed on the n-th note Y_n can be described with W different features such as $Y_n = (f_0, f_1, \ldots, f_W)$. Assuming that X_n is dependent only on X_{n-1}, the joint probability of $\mathbf{X}_{0:N}$ and $\mathbf{Y}_{0:N}$ can be factorized such as

$$P(\mathbf{X}_{0:N}, \mathbf{Y}_{0:N}) = P(X_0)P(Y_0|X_0)\prod_{n=1}^{N} P(X_n|X_{n-1})P(Y_n|X_n). \quad (1)$$

We assume that S_1 and S_K represent the beginning and end of a gesture respectively, and constraint the state transitions as a quasi-left-right model (Figure 2a). This transition diagram indicates that there would be a gesture boundary on the transition of $\mathrm{S}_K \rightarrow \mathrm{S}_1$.

Y_n should be defined with features describing its role for the gesture implementation. Therefore, we define Y_n with a 2-dim. vector containing relative value changes on $(n-1, n)$ and $(n, n+1)$ such as

$$Y_n^{\mathrm{Interval}} = (\mathrm{Interval}_{(n,n-1)}, \mathrm{Interval}_{(n+1,n)}), \quad (2)$$

$$Y_n^{\mathrm{Rhythm}} = (\frac{\mathrm{Duration}_n}{\mathrm{Duration}_{n-1}}, \frac{\mathrm{Duration}_{n+1}}{\mathrm{Duration}_n}), \quad (3)$$

$$Y_n^{\mathrm{Tempo}} = (\frac{\mathrm{Bpm}_n}{\mathrm{Bpm}_{n-1}}, \frac{\mathrm{Bpm}_{n+1}}{\mathrm{Bpm}_n}), \quad (4)$$

$$Y_n^{\mathrm{Loudness}} = (\frac{\mathrm{Velocity}_n}{\mathrm{Velocity}_{n-1}}, \frac{\mathrm{Velocity}_{n+1}}{\mathrm{Velocity}_n}). \quad (5)$$

Quantising tempo and loudness values with a number of steps L, we can model Equation 1 with a discrete Hidden Markov Model, which is one of the simplest statistical latent state models.

2.2 Unsupervised Learning

Modelling gestures in a given performance with HMMs is equivalent to finding the initial state probabilities π, the state transition probabilities A and the state emission probabilities B from the input performance data, which satisfy the constraints on the state transitions. The standard Baum-Welch Algorithm provides an iterative solution for a Maximum Likelihood Estimation of those model parameters. Since the algorithm is based on the Expectation-Maximization approach, this unsupervised learning process can be understood as a clustering of observed property values into latent states while considering state transitions.

Once we have trained all model parameters, we can estimate the sequence of latent states given a property sequence by computing

$$\mathbf{X}^*_{0:N} = \arg\max_{\mathbf{X}_{0:N}} P(\mathbf{X}_{0:N}|\mathbf{Y}_{0:N}; \pi, A, B), \quad (6)$$

and this can be efficiently calculated with the standard Viterbi Algorithm.

The number of states K is related to the length of a gesture. If K is small, the model tends to estimate gestures as smaller groups, and if K is large, larger groups will be estimated as gestures. Considering that the model likelihood indicates how good the model explains the patterns observed in the given performance, the information criteria such as log-likelihood, Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC) can be applied to determine the number of states K.

2.3 Effectiveness Measure

Each property has a different effect on the gesture identification, and we focus on the variations of each property patterns to measure its relevance. If the variations are randomly distributed, it is difficult to find particular gestures in the sequence, and the state emission probabilities would be distributed uniformly. If there are clear repetitions of a particular gesture, then we can identify it easily, and the kurtosis of each state emission probability distribution would be high. This suggests an entropy-based measure C such as,

$$C(B^{\mathcal{M}}) = \frac{1}{K}\sum_{k=1}^{K} H(p^B_{\mathrm{S}_k}), \quad (7)$$

where $B^{\mathcal{M}}$ and $H(p^B_{S_k})$ is the state emission probability matrix of model $\mathcal{M}$, and the entropy of each emission probability distribution, respectively. With this effectiveness measure we can then select particular property gesture models as shown in Figure 2b.

The concept can be understood as a statistical multi-stream approach, since we assume that all properties were generated from individual information sources. Unlike factorial HMMs, this allows a factorisation of not only the latent states, but also the observation sequences.

3. RESULTS

3.1 Gesture Detection

We tested our model on some recorded performances selected from the CrestMuse PEDB[1]. Figure 3a shows a part of the gesture analysis result on Schumann's *Träumerei* performed by I. Hemming. In this piece we can find repetitions of a rhythmic pattern in bars 1-2, 5-6, 9-10 and 13-14, and variations in bars 3-4, 7-8 and so forth. The brackets in bold under the score indicate rhythmic patterns estimated with our model, which largely matched the motivic elements illustrated with grey boxes. There are also pattern variations in pitch, but the rhythmic patterns seem to be more obvious than those in pitch, and this is indicated by a smaller value of the effectiveness measure as shown in Figure 3d.

Looking at tempo changes, we can find a pattern slowed down on certain positions of the composition. In particular, we observe characteristic tempo changes in bars 1-2, 5-6 and so forth. However, such tempo pattern boundaries are often not coinciding with the rhythmic ones. The up-ward arrows indicate the tempo pattern boundaries estimated with our model, and the characteristic tempo changes were largely detected.The loudness changes, which are not depicted in Figure 3a, had less obvious patterns, and its effectiveness measure was indeed higher than the tempo property model's one.

Figures 3b and 3c illustrate the state emission probabilities of rhythm and tempo property models. The rhythmic patterns observed in Figure 3 start with a quarter note, continue with a sequence of 8th notes, and end with a half note. The emission probabilities captured such changes and clustered them into latent states. For example, the most probable rhythm property under S_K had a feature such that the duration of the n-th note is four times longer than the $(n-1)$-th one. Similarly, the tempo model captured a slowed-down tempo pattern observable in the given performance.

Hence, our model could detect patterns observed in different properties, and detect their boundaries using both symbolic notation and expressive properties measured in a recorded performance. In addition, the entropy-based measure was used to estimate the importance of each property for the gesture identification.

3.2 Different Interpretations

Figure 4 shows different interpretations of the same piece rendered with diverse gestures. The performance illustrated in Figure 4a has typical "arch"-like loudness changes in the first 4 bars. Another performance illustrated in Figure 4b also has softened loudness in the 4th measure, but very similar loudness patterns are found in bars 1-2 and 5-6, which support the rhythmic pattern repetitions. On the other hand, we can observe different interpretations in bars 7-8. In Figure 4b we can find a pattern that softens loudness over both measures. In Figure 4a we can observe such a pattern twice, one until the second and half beat of the 7th measure, and another until the third beat of the 8th measure. Those different loudness patterns convey different gestures.

The arrows indicate the loudness pattern boundaries estimated with our model, and they were different in each performance since the two input performances have different gestural structures. In particular, the additional loudness softening in bars 7-8 of 4a makes a difference between two interpretations, and those different gestures were identified by the model.

[1]CrestMuse PEDB, http://www.crestmuse.jp/pedb

Subset	Region	Count	Precision	Recall	F-Score	TPR
erk	GER	1700	0.54	0.48	0.51	0.18
kinder	GER	213	0.54	0.43	0.47	0.21
han	CHN	1187	0.33	0.43	0.37	0.15

Table 1: Test result for different subsets of the ESAC-DB.

3.3 Results For A Larger Corpus

To evaluate how successful the model is in identifying the gestural structure of music, the ESAC[2] database was used as a larger corpus. It includes subsets for different countries and cultural regions, and many of them have phrase boundary annotations cross-checked by experts. Since it contains compositional properties only, we tried to detect boundaries only based on pitch and rhythm for folksongs found in 3 different subsets, namely *erk*, *kinder* and *han*. The number of states for each property gesture model was determined based on the model log-likelihood, because it outperformed other information criteria[3]. We selected then one of the pitch and rhythm gesture HMMs based on the effectiveness measures discussed above, and estimated boundaries with the selected model. We compared the estimations to the human annotations, and calculated Precision, Recall and F-score on each subset [8]. Precision measures how many boundaries were matched among the estimated ones, and Recall measures how many boundaries were detected from the annotated ones. F-Score is their harmonic mean. In addition, we calculated the Time Precision Rate (TPR) by computing average deviation (error) of the estimated boundaries from the annotated ones, and divided it by the average number of notes of the estimated gestures.

Table 1 shows the results for each subset. The F-scores for the *erk*, *kinder* and *han* subsets were 0.51, 0.47 and 0.37. The average TPR of three subsets was 0.18. Comparing with other methods, the score is still not high [5]. However, a half of the manually annotated gestures could be correctly identified by the model, and considering that the model is based on a simple discrete HMM, we can expect an improvement with a more complex model to capture complicated patterns.

4. DISCUSSION

Test results indicate that the model could identify gestures in a music performance by detecting patterns of different properties in a statistical way, and the entropy-based measure could estimate the relevance of each property on the gesture identification. Since the statistical model learns patterns in the given performance automatically, the approach

[2]Essen Associative Code and Folksong Database, http://www.esac-data.org

[3]F1-scores on the *kinder* subset of the ESAC with log-likelihood, AIC and BIC were 0.47, 0.43 and 0.42.

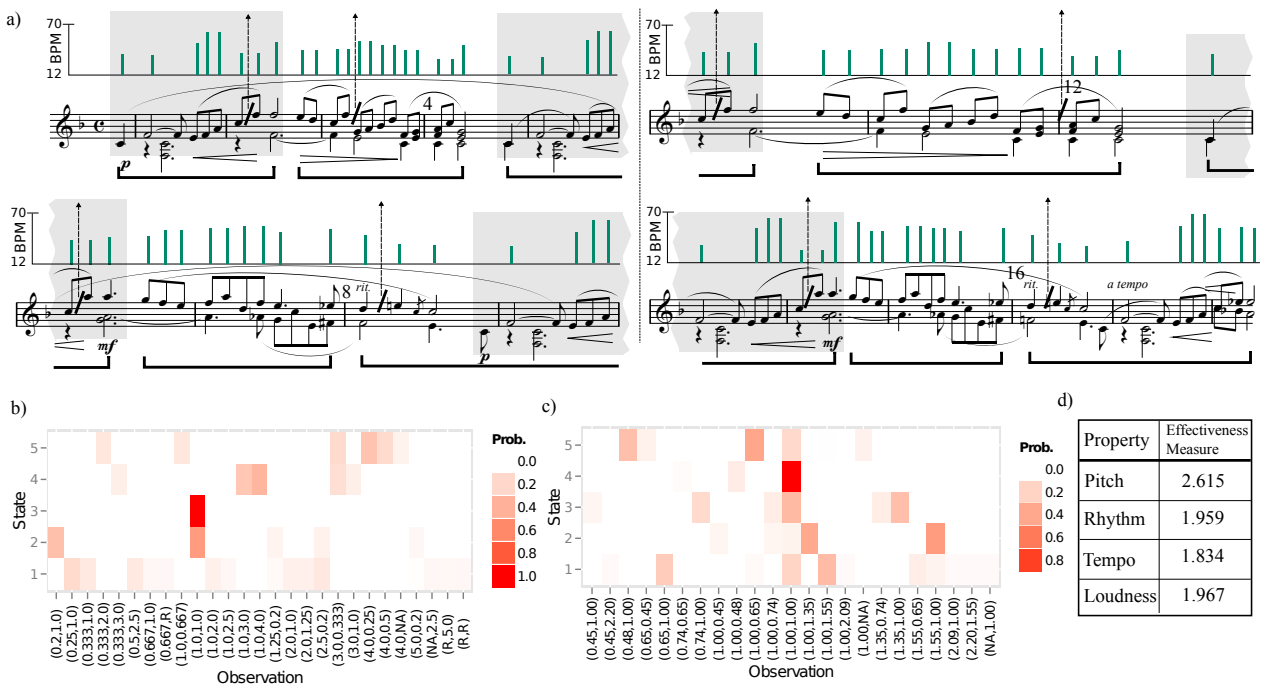

Figure 3: a) Gesture analysis result on R. Schumann's *Träumerei* performed by I. Hemming found in CrestMuse PEDB (ID: sch-kdz007-hemmi-y). b) State emission probabilities of rhythm property model. c) State emission probabilities of tempo property model. d) Effectiveness measures for each property models. The number of states K was 5, and the number of quantisation steps L was 4.

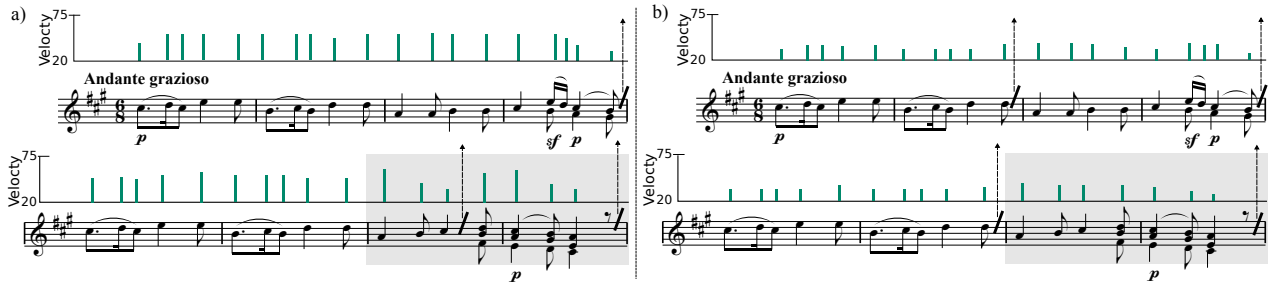

Figure 4: a) W.A. Mozart, Piano Sonata KV331-1, the first 8 bars performed by H. Nakamura (ID: moz-snt331-1-nakam-g) b) Same piece performed by N. Shimizu (ID: moz-snt331-1-shimi-g). The number of states K was 5, and the number of quantisation steps L was 4.

could be applicable to a wide range of music interpretations, even in different cultures.

Due to the simplicity of the gesture HMM, however, a complex gestural structure with different kinds of patterns is still difficult to model. In addition, the effectiveness measure ,as it has been used here, assumes that one property dominates the gestural structure. In a more complex model, this assumption could, of course, be relaxed. To this end, a more generalised statistical modelling framework such as Dynamic Bayesian Network could be suitable.

5. REFERENCES

[1] E. Cambouropoulos. The Local Boundary Detection Model (LBDM) and its application in the study of expressive timing. In *Proc. of ICMC*, 2001.

[2] E. Cheng and et al. A local maximum phrase detection method and the analysis of phrasing strategies in expressive performances. In *Proc. of Math and Computation in Music*, 2007.

[3] R. Hatten. *Interpreting Musical Gestures, Topics, and Tropes: Mozart, Beethoven, Schubert (Musical Meaning and Interpretation)*. Indiana University Press, 2004.

[4] M. T. Pearce and et al. A comparison of statistical and rule-based models of melodic segmentation. In *Proc. of ISMIR*, 2008.

[5] M. Pierce and et al. A comparison of statistical and rule-based models of melodic segmentation. In *Proc. of ISMIR*, 2008.

[6] J. Rink and et al. Motive, gesture, and the analysis of performance. In *New Perspectives on Music and Gesture*, pages 267–292. Ashgate, 2011.

[7] D. Temperley. *Music and Probability*. MIT Press, 2007.

[8] C. van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, 1979.