Polyhymnia: An automatic piano performance system with statistical modeling of polyphonic expression and musical symbol interpretation

Tae Hun Kim, Satoru Fukayama, Takuya Nishimoto and Shigeki Sagayama
Graduate School of Information Science and Technology
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan
{kim, fukayama, nishi, sagayama}@hil.t.u-tokyo.ac.jp

ABSTRACT

We developed an automatic piano performance system called Polyhymnia that is able to generate expressive polyphonic piano performances with music scores so that it can be used as a computer-based tool for an expressive performance. The system automatically renders expressive piano music by means of automatic musical symbol interpretation and statistical models of structure-expression relations regarding polyphonic features of piano performance. Experimental results indicate that the generated performances of various piano pieces with diverse trained models had polyphonic expression and sounded expressively. In addition, the models trained with different performance styles reflected the styles observed in the training performances, and they were well distinguishable by human listeners. Polyhymnia won the first prize in the autonomous section of the Performance Rendering Contest for Computer Systems (Rencon) 2010.

Keywords

performance rendering, polyphonic expression, statistical modeling, conditional random fields

1. INTRODUCTION

We developed an automatic piano performance system called Polyhymnia. To our knowledge, it is the first system that is able to learn and predict polyphonic expression in piano music with diverse performance styles. Human prefer an expressive performance rather than a flat performance obtained by direct converting into MIDI format, and therefore computer-based tools for an expressive music performance would be useful for computer-aided music creations and performances. Unfortunately, automatic rendering of an expressive performance with a music score is a very difficult problem since expressive performance is one of the most complicated human tasks, and its mechanism is still not clear.

There exist many instruments for performing music. Since each instrument has different mechanical design, developing an universal method for automatic renditions of any musical instruments is extremely difficult. We are focusing on piano renditions since piano has abundant solo pieces so that it promises a useful application for computer-aided music creations and performances. Fortunately, musical expression in piano music can be represented with only 3 expression parameters: instantaneous tempo, loudness (velocity) and performed duration. Such a simple parametric representation allows us to develop a simple model of piano performance that can be well encoded in MIDI format.

Polyhymnia fully automates an expressive piano performance. Musical symbols provide a basic guideline for an expressive performance and they can be interpreted in several ways. Therefore we propose flexible parametric models for their automatic interpretation. Polyphonic features of expressive piano performance is very important since piano music is usually polyphonic. We discuss them in this paper and call musical expression with such features *polyphonic expression*. We proposed a statistical modeling of polyphonic piano renditions and showed that generated performances with polyphonic expression sounded better than performances without it [4]. We briefly describe the idea behind the proposed modeling and show how to implement it with Conditional Random Fields (CRFs).

An automatic piano performance system should be able to deal with various unknown piano pieces. Experimental results on performances generated by Polyhymnia with various compositions indicate that they had polyphonic expression and sounded expressively. A piano piece can be performed with diverse performance styles. One of the benefits of the proposed modeling is that diverse models can be easily obtained by training with various performance styles. Experimental results on diverse performances generated by Polyhymnia indicate that each trained model reflected the style observed in the training performance set.

Polyhymnia participated in the Performance Rendering Contest for Computer Systems (Rencon) 2010, and won the first prize in the autonomous section of the contest.

2. RELATED WORK

Several systems for automatic piano renditions are proposed [5]. Director Musices and the Rubato system utilize sets of performance rules extracted by music experts. Kagurame series and COPER are based on several searching algorithms from human performances. ESP, YQX and Usapi try to statistically model musical expression in piano music, whose parameters are learned from training performances. Most of those systems discuss renditions of monophonic melodies, and polyphonic renditions have not been well discussed due to computational complexity and necessity of a huge amount of data. In addition, automatic interpretation of musical symbols were not well discussed since they input a score in MIDI-like format, and it is based on very simple rules.

NIME'11, 30 May–1 June 2011, Oslo, Norway.

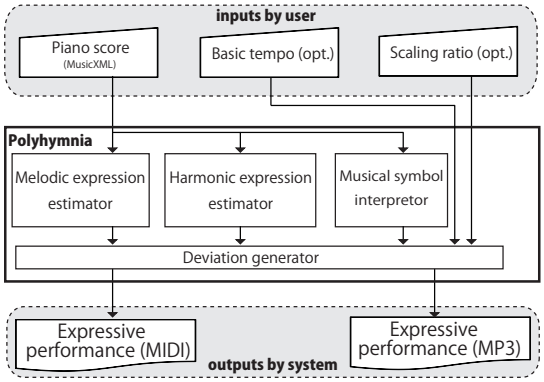

Figure 1: Polyhymnia architecture.

Some commercial notation softwares are also able to provide an expressive performance of a given piece. Although it is unclear how they generate musical expression, their methods are probably based on interpretation of musical symbols with simple rules.

3. SYSTEM OVERVIEW

To obtain an expressive piano performance with Polyhymnia, users requires only to input an piano score in MusicXML[1] format without any other configurations. Unlike MIDI format, MusicXML is able to encode almost all kinds of musical symbols digitally. Encoded musical symbols are automatically interpreted with parametric models that are flexible to generate various interpretations of each occurrence of a symbol. Polyphonic expression is learned and generated with Conditional Random Fields for polyphonic piano renditions. The depth of generated musical expression can be controlled by scaling ratios. The system provides expressive performances in MP3 and MIDI formats (Figure 1).

4. MUSICAL SYMBOL INTERPRETATION

4.1 Expression marks

Dynamic marks such as *p*, *mf* and so on, should be mapped to concrete MIDI velocity values. To find such mapping, 15 performances of V. Ashkenazy in CrestMuse PEDB [2] were analyzed. Table 1 shows the analytical result indicating that interpretation of each occurrence of a mark is distributed, and its interpretations in upper and lower staffs are different over all dynamic marks, for example, marks in lower staff are performed softer than those in upper staff. In order to interpret dynamic marks automatically, given marks should be mapped to concrete values with various maps. As a simple solution, Polyhymnia simply maps given marks to the estimated mean values for upper and lower staffs, respectively. However, this can be improved by proper selection of a map for each occurrence of a mark.

crescendo, *diminuendo* and *ritardando* should be interpreted with gradual changes of loudness and tempo. It is well known that human perceives them by exponential changes of sound energy and tempo in BPM. With analysis of human performances, we found that human performers performs such changes in various forms, and interpret *ritardando* with gradual decreasing tempo and loudness. To model such interpretations, we propose an parametric mathematical model for loudness and tempo changes. Let d_t be loudness in MIDI-velocity[2] or instantaneous tempo in log-

[1]http://www.recordare.com/musicxml

[2]MIDI-velocity can be regarded as a logarithmic scale for sound energy.

Table 1: Human interpretation of dynamic marks. Note that all averages and standard deviations are in MIDI velocity. *ppp* and *mp* were not occurred in the data.

Upper staff

	ppp	*pp*	*p*	*mp*	*mf*	*f*	*ff*	*fff*
Occur.	-	157	2087	-	67	1490	418	19
Avg.	-	50	52	-	58	67	76	98
St. dev.	-	14	15	-	8	15	16	2

Lower staff

	ppp	*pp*	*p*	*mp*	*mf*	*f*	*ff*	*fff*
Occur.	-	150	3169	-	53	1538	353	12
Avg.	-	37	37	-	47	57	73	101
St. dev.	-	13	11	-	9	20	19	11

BPM at time t. Then, its gradual changes over t can be modeled as

$$d_t = d_0(\beta \cdot t^{\alpha} + 1), \tag{1}$$

where d_0 is start value, β is the parameter for expression depths and α is the parameter for shapes. If α is 1.0, energy and tempo in BPM are changing exponentially. With different setting of α and β, each occurrence of a mark can be interpreted in various forms. As a simple solution, Polyhymnia interpret all occurrences of a mark with fixed parameter values. However, this can be improved by automatic determination of parameter values for each occurrence of a mark.

4.2 Ornaments

Mordent, *turn*, *trill* and grace notes are performed with additional notes. Since such additional notes decorate their parent notes, we can assume that their loudness is determined based on their parent note's loudness. However, human is not able to perform a note sequence with a constant velocity. Assuming that such motor error is following Gaussian distribution, loudness of the ith additional note d_i can be modeled as

$$d_i = d_0 + \mathcal{N}(0, \sigma^2), \tag{2}$$

where d_0 is the loudness of the parent note. σ^2 controls fluctuation ranges of loudness.

Arpeggio indicates that onset time of each arpeggiated note should be delayed one after another. Since human is not able to perform such notes with a constant delay, we can assume that it contains Gaussian noise. Then, onset time of ith arpeggiated note d_i can be modeled as

$$d_i = d_0 + i \cdot \Delta + \mathcal{N}(0, \sigma^2), \tag{3}$$

where Δ is a delay time, and d_0 is the onset time of the lowest arpeggiated note.

5. STATISTICAL MODELING OF POLYPHONIC PIANO RENDITIONS

5.1 Polyphonic expression in piano music

Although musical symbols provides a basic guideline for an expressive performance, musical expression in piano music is much more complicated, for example, instantaneous tempo, loudness and performed duration are fluctuating over time, even if there are no musical symbols for them. In addition, an expressive piano performance has polyphonic expression whose features include:

- Each voice expression has fluctuations of loudness and performed duration over time, and it is not always same to the other voices.

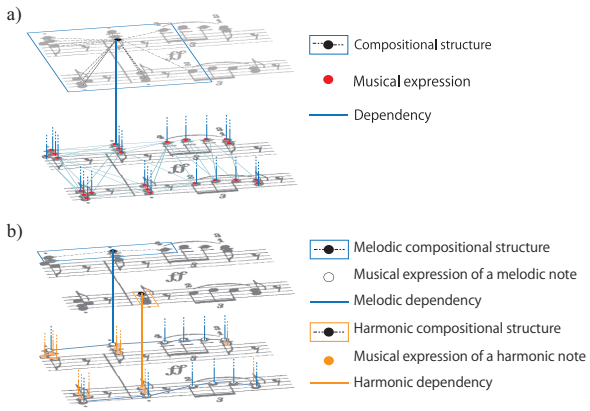

Figure 2: Complex dependency networks of polyphonic expression in piano music (a). Simplified dependency networks by introducing melodic and harmonic dependencies (b).

- Expression of each note in a chord is not always same to the other notes in the chords. Playing a chord with different combinations of note expression results different sounds of the chord.

In order to learn and predict polyphonic expression in piano music, we proposed a statistical modeling of polyphonic piano renditions and showed its efficiency for improving a machine-rendered piano performance [4]. In following subsections, we briefly describe the idea behind the modeling and show how to implement it with Conditional Random Fields.

5.2 Probabilistic formulation

Prediction of an expressive performance D given a piano score $\mathbf{S}$ can be formulated probabilistically such as

$$\hat{D} = \arg\max_{D} P(D|\mathbf{S};\Theta), \tag{4}$$

where Θ is the parameters of the distribution. To model $P(D|\mathbf{S};\Theta)$, we assume that a note expression is dependent on its compositional structure represented with score features and on the other note expressions. Figure 2a shows a dependency network of polyphonic piano music. In case of polyphonic expression, such dependency is very complex, and therefore it is hard to model it with computational tractability, and a huge amount of training data is necessary for learning model parameters Θ. Therefore, an approximation to polyphonic expression is necessary for a tractable modeling.

To simplify dependency in polyphonic renditions, we proposed an approximation with melodic and harmonic dependencies. Figure 2b shows an example of simplified dependency network with the proposed approximation. We believe that such approximation promises a perceptually best performance. This is because human perceives

- different voice expressions sounding simultaneously,
- different sounds of a given harmony,
- expressions of outer voices easier than that of inner voices [3].

Hence, $P(D|\mathbf{S};\Theta)$ can be approximated such as

$$P(D|\mathbf{S}) = P(D^{\mathrm{m}^u}|\mathbf{S}^{\mathrm{m}^u}) \cdot P(D^{\mathrm{m}^l}|\mathbf{S}^{\mathrm{m}^l}) \cdot \prod_{h^u=1}^{H^u} P(D^{h^u}|\mathbf{S}^{h^u}) \cdot \prod_{h^l=1}^{H^l} P(D^{h^l}|\mathbf{S}^{h^l}), \tag{5}$$

where $P(D^{\mathrm{m}^u}|\mathbf{S}^{\mathrm{m}^u})$ and $P(D^{\mathrm{m}^l}|\mathbf{S}^{\mathrm{m}^l})$ are distributions of melodic expression in the uppermost and lowermost voices, respectively, and $P(D^{h^u}|\mathbf{S}^{h^u})$ and $P(D^{h^l}|\mathbf{S}^{h^l})$ are distributions of harmonic expressions in upper and lower staffs, respectively.

Since such approximation allows Markov assumption on both of melodic and harmonic dependencies, they can be modeled with statistical models with hidden state transitions, such as Dynamic Bayesian Networks, Hidden Markov Models and Conditional Random Fields. Considering that our goal is to estimate a note expression sequence given a sequence of score feature vectors representing a piano score, we believe that CRF is one of the best frameworks for modeling those dependencies.

5.3 Modeling with Conditional Random Fields

We assume that a melodic compositional structure is represented with score features, such as pitch, duration, note interval and so on, and a harmonic compositional structures is represented with score features, such as pitch, duration and so on. Also, we assume that melodic expression is represented with instantaneous tempo, loudness and performed duration, and harmonic expression is represented with onset time differences, loudness and performed duration[3].

Although score features and expression parameters of melodic and harmonic dependencies are different to each other, they can be modeled with CRFs with the same model structure. Let d_n and D be the nth melodic or harmonic expression and its sequence, respectively. Let $\mathbf{S}$ be a sequence of score feature vectors representing melodic or harmonic compositional structures, and s_k be the kth score feature. Assuming that d_n is only dependent on d_{n-1} (Markov assumption), we can define the jth feature function F_j such as

$$F_j(D,\mathbf{S}) = \sum_{n=1}^{N} \delta(\{d_{n-1}, d_n, s_k\}_j, n), \tag{6}$$

where $\delta(\cdot)$ returns 1, if the jth triple from all possible triples of $\{d_{n-1}, d_n, s_k\}$ is occurred at position n, and 0, otherwise.

Introducing a weight variable θ_j for each F_j and according to the Maximum Entropy Principle, $P(D|\mathbf{S};\Theta)$ can be defined such as

$$P(D|\mathbf{S};\Theta) = \frac{1}{Z(\mathbf{S},\Theta)} \exp \sum_j \theta_j F_j(D,\mathbf{S}), \tag{7}$$

where

$$Z(\mathbf{S},\Theta) = \sum_{D'} \exp \sum_j \theta_j F(D',\mathbf{S}). \tag{8}$$

Model parameters Θ can be learned from training performances with Maximum Likelihood Estimation by an iterative algorithm, such as Stochastic Gradient Descent [1]. Once Θ is estimated, we can predict an expressive performance with equation (4), and this can be efficiently computed with Forward-backward algorithm and Dynamic Programming technique [6].

6. EXPERIMENTAL EVALUATION

6.1 Generation quality

An automatic music performance system should be able to render *unknown* pieces in various compositional styles. In order to evaluate Polyhymnia in this aspect, piano pieces in various compositional styles were rendered by the system, and evaluated by 19 human listeners[4]. We rendered

[3]Details of the melodic and harmonic score features and the expression parameters can be found in [4]

[4]2 professional musicians, 13 hobby musicians and 2 non-musicians participated in the listening experiments.

Table 2: Test pieces used in the experiment.

ID	Composer	Piece	Tempo
CF	Chopin	Mazurka no. 5, op. 7-1	fast
CS	Chopin	Sonata no. 2-3, op. 35	slow
MF	Mozart	Sonatina no.5-3, KV. 439	fast
MS	Mozart	Marche Funebre, KV. 453a	slow
RT	S. Joplin	The Entertainer (ragtime)	middle
GR	Grieg	7 Lyric Piec., 7. Rem., op. 71	slow

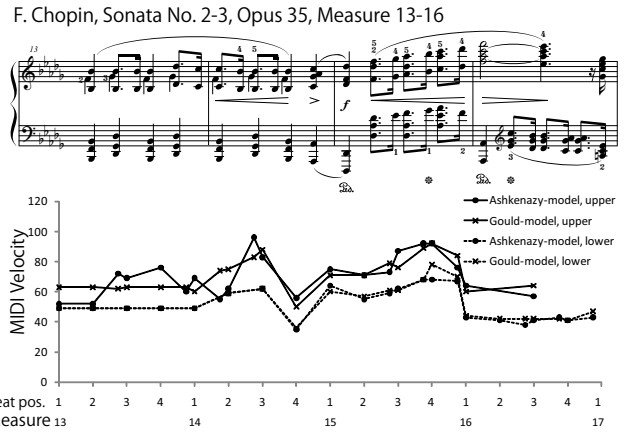

Figure 3: Generated performances by Polyhymnia: F. Chopin, Sonata no. 2-3, op. 35.

6 unknown pieces with Polyhymnia as shown in Table 2. Note that the test pieces included not only F. Chopin and W. A. Mozart's pieces, but also E. Grieg and S. Joplin's pieces whose compositional styles are quite different to the training pieces.

In order to render expressive performances with different performance styles, we prepared two different models such as *Ashkenazy-model* and *Gould-model* trained with 15 performances of V. Ashkenazy[5] and 7 performances of G. Gould[6], respectively (CrestMuse PEDB). We prepared 5 performances for each piece such as performance without expression (D), by musical symbol interpretation only (E), generated with *Ashkenazy-model* (S1), generated with *Gould-model* (S2) and by a human performer (H). All of those sound samples were blind to the listeners, and their human-likeness and musicality were evaluated using 6-level-scales.

Figure 3 shows an example of generated performances by Polyhymnia. The results indicate that the had polyphonic expression, and their fluctuations were different to each other. Figure 4a shows the average scores of the 6 test pieces. Analysis of Variance on those average differences with $p < 0.05$ indicate that performances generated by Polyhymnia sounded better than performances without expression. Score difference between S1 and H was not significant. This means that performances generated with *Ashkenazy-model* sounded expressively like human performances do. Score differences between S2 and H were not significant in some particular pieces. This means that some performances generated with *Gould-model* sounded expressively like human performances do.

[5]Prelude no. 1, 4, 7, 15, 20, Etude op. 10-3, 10-4, 25-11, Waltz op. 18, 34-2, 64-2, 69-1, 69-2, Nocturne no. 2, 10.
[6]Piano Sonata KV279-1, 279-2, 279-3, 331-1, 545-1, 545-2, 545-3.

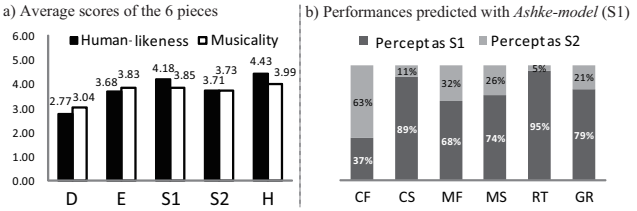

Figure 4: Average scores of the 6 pieces (a). Style classification result of the 6 S1 (b).

6.2 Subjective style identification

In order to know if each trained model reflected the style observed in the training data, we conducted an another listening experiment for subjective style identification. 3 piano pieces, which were not included in the test pieces, were generated with both trained models (total 6 performances) and the participants listened to them to remember the style each model reflected. After that, the participants listened to the 12 S1 and S2 blind in a random order.

Figure 4b shows the style identification result of the 6 S1. The result shows that 5 out of 6 pieces were well identified by the listeners, and the average identification rate was 73.6%. The identification result of the 6 S2 was similar, and the average identification rate was 73.6%. Those results indicate that each trained model reflected the style observed in training data, and those styles were perceptually distinguishable by human listeners.

7. CONCLUSION

We introduced an automatic piano performance system called Polyhymnia that is able to learn and predict polyphonic expression, and interpret musical symbols automatically. Experimental evaluations on generated performances indicate that diverse performances of various compositions generated by the system had polyphonic expression and sounded expressively, and their performance styles were perceptually well distinguishable by human listeners.

We believe that modeling hierarchical structures of a given piece would improve a machine-rendered piano performance. By introducing additional model parameters controlled by users through an interface, Polyhymnia can be extended to an interactive music performance system.

8. REFERENCES

[1] L. Bottou. Stochastic gradient learning in neural networks. In *Proc. Neuro-Nimes*, Nimes, France, 1991. EC2.

[2] M. Hashida and et al. A new database describing deviation information of performance expressions. In *Proc. ISMIR*, pp. 489–494, 2008.

[3] D. Huron and et al. The avoidance of inner-voice entries: perceptual evidence and musical practice. *Music Perception*, 7(1):43–48, 1989.

[4] T. H. Kim and et al. Performance rendering for polyphonic piano music with a combination of probabilistic models for melody and harmony. In *Proc. SMC*, pp. 23–30, 2010.

[5] A. Kirke and et al. A survey of computer systems for expressive music performance. *ACM Comput. Surv.*, 42(1), 2009.

[6] J. Lafferty and et al. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proc. ICML*, pp. 282–289, 2001.