MERGED-OUTPUT HMM FOR PIANO FINGERING OF BOTH HANDS

Eita Nakamura
National Institute of Informatics
Tokyo 101-8430, Japan
eita.nakamura@gmail.com

Nobutaka Ono
National Institute of Informatics
Tokyo 101-8430, Japan
onono@nii.ac.jp

Shigeki Sagayama
Meiji University
Tokyo 164-8525, Japan
sagayama@meiji.ac.jp

ABSTRACT

This paper discusses a piano fingering model for both hands and its applications. One of our motivations behind the study is automating piano reduction from ensemble scores. For this, quantifying the difficulty of piano performance is important where a fingering model of both hands should be relevant. Such a fingering model is proposed that is based on merged-output hidden Markov model and can be applied to scores in which the voice part for each hand is not indicated. The model is applied for decision of fingering for both hands and voice-part separation, automation of which is itself of great use and were previously difficult. A measure of difficulty of performance based on the fingering model is also proposed and yields reasonable results.

1. INTRODUCTION

Music arrangement is one of the most important musical activities, and its automation certainly has attractive applications. One common form is piano arrangement of ensemble scores, whose purposes are, among others, to enable pianists to enjoy a wider variety of pieces and to accompany other instruments by substituting the role of orchestra. While certain piano reductions have high technicality and musicality as in the examples by Liszt [8], those for vocal scores of operas and reduction scores of orchestra accompaniments are often faithful to the original scores in most parts. The most faithful reduction score is obtained by gathering every note in the original score, but the result can be too difficult to perform, and arrangement such as deleting notes is often in order.

In general, the difficulty of a reduction score can be reduced by arrangement, but then the fidelity also decreases. If one can quantify the performance difficulty and the fidelity to the original score, the problem of "minimal" piano reduction can be considered as an optimization problem of the fidelity given constraints on the performance difficulty. A method for guitar arrangement based on probabilistic model with a similar formalization is proposed in Ref. [5]. This paper is a step toward a realization of piano reduction algorithm based on the formalization.

The playability of piano passages is discussed in Refs. [3, 2] in connection with automatic piano arrangement. There, constraints such as the maximal number of notes in each hand, the maximal interval being played, say, 10th, and the minimal time interval of a repeated note are considered. Although these constraints are simple and effective to some extent, the actual situation is more complicated as manifested in the fact that, for example, the playability can change with tempos and players can arpeggiate chords that cannot be played simultaneously. In addition, the playability can depend on the technical level of players [3]. Given these problems, it seems appropriate to consider performance difficulty that takes values in a range.

There are various measures and causes of performance difficulty including player's movements and notational complexity of the score [12, 1, 15]. Here we focus on the difficulty of player's movements, particularly piano fingering, which is presumably one of the most important factors. The difficulty of fingering is closely related to the decision of fingering [4, 7, 13, 16]. Given the current situation that a method of determining the fingering costs from first principles is not established, however, it is also effective to take a statistical approach, and consider the naturalness of fingering in terms of probability obtained from actual fingering data. With a statistical model of fingering, the most natural fingering can be determined, and one can quantify the difficulty of fingering in terms of naturalness. This will be explained in Secs. 2 and 3. The practical importance of piano fingering and its applications are discussed in Ref. [17].

Since voice parts played by both hands are not a priori separated or indicated in the original ensemble score, a fingering model must be applicable in such a situation. Thus, a fingering model for both hands and an algorithm to separate voice parts are necessary. We propose such a model and an algorithm based on merged-output hidden Markov model (HMM), which is suited for modeling multi-voice-part structured phenomena [10, 11]. Since multi-voice-part structure of music is common and voice-part separation can be applied for a wide range of information processing, the results are itself of great importance.

2. MODEL FOR PIANO FINGERING FOR BOTH HANDS

2.1 Model for one hand

Before discussing the piano fingering model for both hands, let us discuss the fingering model for one hand. Piano

fingering models and algorithms for decision of fingering have been studied in Refs. [13, 16, 4, 18, 19, 20, 7]. Here we extend the model in Ref. [19] to including chords.

Piano fingering for one hand, say, the right hand, is indicated by associating a finger number $f_n = 1, \cdots, 5$ ($1 =$ thumb, $2 =$ the index finger, $\cdots$, $5 =$ the little finger) to each note p_n in a score [1], where $n = 1, \cdots, N$ indexes notes in the score and N is the number of notes. We consider the probability of a fingering sequence $f_{1:N} = (f_n)_{n=1}^N$ given a score, or a pitch sequence, $p_{1:N} = (p_n)_{n=1}^N$, which is written as $P(f_{1:N}|p_{1:N})$. As explained in detail in Sec. 3.1, an algorithm for fingering decision can be obtained by estimating the most probable candidate $\hat{f}_{1:N} = \operatorname{argmax}_{f_{1:N}} P(f_{1:N}|p_{1:N})$. The fingering of a particular note is more influenced by neighboring notes than notes that are far away in score position. Dependence on neighboring notes is most simply described by that on adjacent notes, and it can be incorporated with a Markov model. It also has advantages in efficiency in maximizing probability and setting model parameters. Although the probability of fingering may depend on inter-onset intervals between notes, the dependence is not considered here for simplicity.

As proposed in Ref. [18, 19], the fingering model can be constructed with an HMM. Supposing that notes in score are generated by finger movements and the resulting performed pitches, their probability is represented with the probability that a finger would be used after another finger $P(f_n|f_{n-1})$, and the probability that a pitch would result from succeeding two used fingers. The former is called the transition probability, and the latter output probability. The output probability of pitch depends on the previous pitch in addition to the corresponding used fingers, and it is described with a conditional probability $P(p_n|p_{n-1}, f_{n-1}, f_n)$. In terms of these probabilities, the probability of notes and fingerings is given as

$$P(p_{1:N}, f_{1:N}) = \prod_{n=1}^{N} P(p_n|p_{n-1}, f_{n-1}, f_n) P(f_n|f_{n-1}), \tag{1}$$

where the initial probabilities are written as $P(f_1|f_0) \equiv P(f_1)$ and $P(p_1|p_0, f_0, f_1) \equiv P(p_1|f_1)$. The probability $P(f_{1:N}|p_{1:N})$ can also be given accordingly.

To train the model efficiently, we assume some reasonable constraints on the parameters. First we assume that the probability depends on pitches only through their geometrical positions on the keyboard which is represented as a two-dimensional lattice (Fig. 1). We also assume the translational symmetry in the x-direction and the time inversion symmetry for the output probability. If the coordinate on the keyboard is written as $\boldsymbol{\ell}(p) = (\ell_x(p), \ell_y(p))$, the assumptions mean that the output probability has a form $P(p'|p, f, f') = F(\ell_x(p') - \ell_x(p), \ell_y(p') - \ell_y(p); f, f')$, and it satisfies $F(\ell_x(p') - \ell_x(p), \ell_y(p') - \ell_y(p); f, f') = F(\ell_x(p) - \ell_x(p'), \ell_y(p) - \ell_y(p'); f', f)$. A model for each hand can be obtained in this way, and it is written as $F_\eta(\ell_x(p') - \ell_x(p), \ell_y(p') - \ell_y(p); f, f')$ with $\eta = L, R$.

[1] We do not consider the so-called finger substitution in this paper.

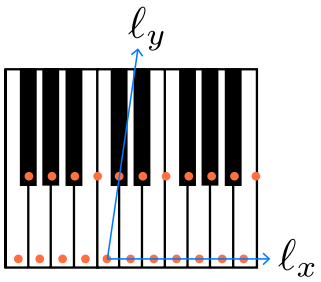


Figure 1. Keyboard lattice. Each key on a keyboard is represented by a point of a two-dimensional lattice.

It is further assumed that these probabilities are related by reflection in the x-direction, which yields $F_L(\ell_x(p') - \ell_x(p), \ell_y(p') - \ell_y(p); f, f') = F_R(\ell_x(p') - \ell_x(p), \ell_y(p') - \ell_y(p); f, f')$.

The above model can be extended to be applied for passages with chords, by converting a polyphonic passage to a monophonic passage by virtually arpeggiating the chords [7]. Here, notes in a chord are ordered from low pitch to high pitch. The parameter values can be obtained from fingering data.

2.2 Model for both hands

Now let us consider the fingering of both hands in the situation that it is unknown a priori which of the notes are to be played by the left or right hand. The problem can be stated as associating the fingering information $(\eta_n, f_n)_{n=1}^N$ for the pitch sequence $p_{1:N}$, where $\eta_n = L, R$ indicates the hand with which the n-th note is played.

One might think to build a model of both hands by simply extending the one-hand model and using (η_n, f_n) as a latent variable. However, this is not an effective model as far as it is a first-order Markov model since, for example, probabilistic constraints between two successive notes by the right hand cannot be directly incorporated when they are interrupted by other notes of the left hand. Using higher-order Markov models leads to the problem of increasing number of parameters that is hard to train as well as the increasing computational cost. The underlying problem is that the model cannot capture the structure of dependencies that is stronger among notes in each hand than those across hands.

Recently an HMM, called merged-output HMM, is proposed that is suited for describing such voice-part-structured phenomena [10, 11]. The basic idea is to construct a model for both hands by starting with two parallel HMMs, called part HMMs, each of which corresponds to the HMM for fingering of each hand, and then merging the outputs of the part HMMs. Assuming that only one of the part HMMs transits and outputs an observed symbol at each time, the state space of the merged-output HMM is given as a triplet $k = (\eta, f_L, f_R)$ of the hand information $\eta = L, R$ and fingerings of both hands: η indicate which of the HMMs transits, and f_L and f_R indicate the current states of the part HMMs. Let the transition and output probabilities

of the part HMMs be $a^{\eta}_{ff'} = P_\eta(f'|f)$ and $b^{\eta}_{ff'}(\boldsymbol{\ell}) = F_\eta(\boldsymbol{\ell}; f, f')$ $(\eta = L, R)$. Then the transition and output probabilities of the merged-output HMM are given as

$$a_{kk'} = \begin{cases} \alpha_L a^L_{f_L f'_L} \delta_{f_R f'_R}, & \eta' = L; \\ \alpha_R a^R_{f_R f'_R} \delta_{f_L f'_L}, & \eta' = R, \end{cases} \tag{2}$$

$$b_{kk'}(\boldsymbol{\ell}) = \begin{cases} b^L_{f_L f'_L}(\boldsymbol{\ell}), & \eta' = L; \\ b^R_{f_R f'_R}(\boldsymbol{\ell}), & \eta' = R, \end{cases} \tag{3}$$

where δ denotes Kronecker's delta. Here, $\alpha_{L,R}$ represent the probability of choosing which of the hands to play the note, and practically, they satisfy $\alpha_L \sim \alpha_R \sim 1/2$. As shown in Ref. [11], certain interaction factors can be introduced to Eqs. (2) and (3). Although such interactions may be important in the future [14], we confine ourselves to the case of no interactions in this paper for simplicity.

By estimating the most probable sequence $\hat{k}_{1:N}$, both the optimal configuration of hands $\hat{\eta}_{1:N}$, which yields a voice-part separation, and that of fingers $(\hat{f}_L, \hat{f}_R)_{1:N}$ are obtained. For details of inference algorithms and other aspects of merged-output HMM, see Ref. [11].

2.3 Model for voice-part separation

The model explained in the previous section involves both hands and the used hand and fingers are modeled simultaneously. We can alternatively consider the problem of associating fingerings of both hands as first separating voice parts for both hands, and then associating fingerings for notes in each voice part. In this subsection, a simple model that can be used for voice-part separation is given. The model is also based on a simpler merged-output HMM, and it yields more efficient algorithm for voice-part separation.

We consider a merged-output HMM with a hidden state $x = (\eta, p_L, p_R)$, where $\eta = L, R$ indicates the voice part, and $p_{L,R}$ describes the pitch played in each voice part. If the pitch sequence in the score is denoted by $(y_n)_n$, the transition and output probabilities are written as

$$a_{xx'} = \begin{cases} \alpha_L a^L_{p_L p'_L} \delta_{p_R p'_R}, & \eta' = L; \\ \alpha_R a^R_{p_R p'_R} \delta_{p_L p'_L}, & \eta' = R, \end{cases} \tag{4}$$

$$b_x(y) = \delta_{y, p_\eta}. \tag{5}$$

Here the transition probability $a^{L,R}_{pp'}$ describes the pitch sequence in each voice part directly, without any information on fingerings. The corresponding distributions can be obtained from actual data of piano pieces, as shown in Fig. 2.

So far we have considered a model of pitches and horizontal intervals for voice-part separation. The voice-part-separation algorithm can be derived by applying the Viterbi algorithm to the above model. In fact, a voice part in the score played by one hand is also constrained by vertical intervals since it is physically difficult to play a chord containing an interval far wider than a octave by one hand. The constraint on the vertical intervals can also be introduced in terms of probability.

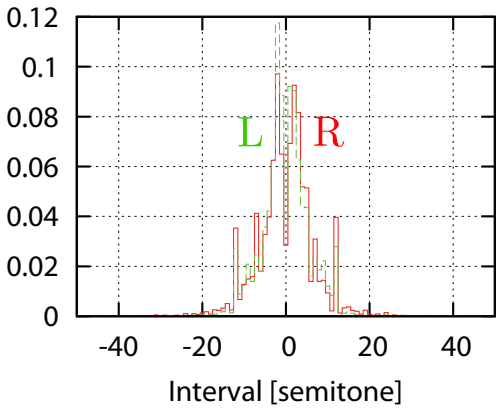


Figure 2. Histograms of pitch transitions in piano scores for each hand.

3. APPLICATIONS OF THE FINGERING MODEL

3.1 Algorithm for decision of fingering

A direct application of the model explained in Secs. 2.1 and 2.2 is the decision of fingering. The algorithm can be derived by applying the Viterbi algorithm. For one hand, the derived algorithm is similar as the one in Ref. [19], but we reevaluated the accuracy since the present model can be applied for polyphonic passages and the details of the models are different.

For evaluation, we prepared manually labeled fingerings of classical piano pieces and compared them to the one estimated with the algorithm. The test pieces were Nos. 1, 2, 3, and 8 of Bach's two-voice inventions, and the introduction and exposition parts from Beethoven's 8th piano sonata in C minor. The training and test of the algorithm was done with the leave-one-out cross validation method for each piece. To avoid zero frequencies in the training, we added a uniform count of 0.1 for every bin.

The averaged accuracy was 56.0% (resp. 55.4%) for the right (resp. left) hand where the number of notes was 5202 (resp. 5539). Since the training data was not big, and we had much higher rate of more than 70% for closed test, the accuracy may improve if a larger set of training data is given. The results were better than the reported values in Ref. [19]. The reason would be that the constraints of the model in the reference was too strong, which is relaxed in the present model. For detailed analysis of the estimation errors, see Ref. [19].

3.2 Voice-part separation

Voice-part separation between two hands can be done with the model described in Sec. 2.3, and the algorithm can be obtained by the Viterbi algorithm. In fact, we can derive a more efficient estimation algorithm which is effectively equivalent since the model has noiseless observations as in Eq. (5).

It is obtained by minimizing the following potential with respect to the variables $\{(\eta_n, h_n)\}, h_n = 0, 1, \cdots, N_h$ for

Table 1. Error rates of the voice-part-separation algorithms.The 0-HMM (resp. 1-HMM, 2-HMM) indicates the algorithm with the zeroth-order (resp. first-order, second-order) HMM.

Pieces	# Notes	0-HMM [%]	1-HMM [%]	2-HMM [%]	Merged-output HMM [%]
Bach (15 pcs)	9638	5.1	5.3	6.1	1.9
Beethoven (2 pcs)	18144	13.0	11.1	11.5	9.28
Chopin (5 pcs)	8508	5.7	4.0	4.29	3.8
Debussy (3 pcs)	3360	17.8	14.8	14.8	18.7
Total	39650	9.9	8.5	8.9	7.1

each note:

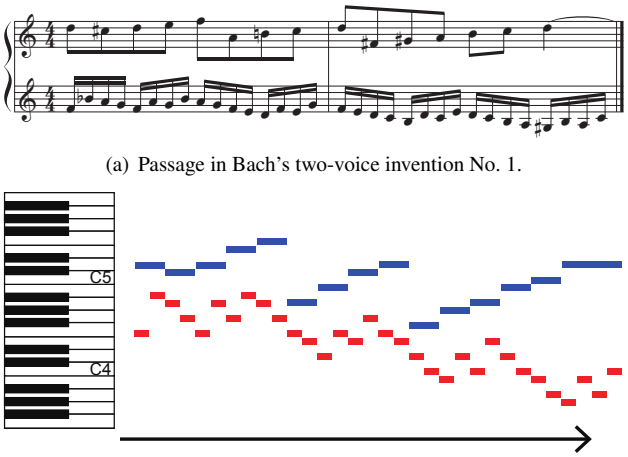
$$V(\boldsymbol{\eta}, \boldsymbol{h}) = -\sum_n \ln Q(\eta_{n-1}, h_{n-1}; \eta_n, h_n), \tag{6}$$

$$Q(\eta_{n-1}, h_{n-1}; \eta_n, h_n) = \begin{cases} \alpha_{\eta_n} a^{(\eta_n)}_{y_{n-1}, y_n} \delta_{h_n, h_{n-1}+1}, & \eta_n = \eta_{n-1}; \\ \alpha_{\eta_n} a^{(\eta_n)}_{y_{n-2-h_{n-1}}, y_n} \delta_{h_n, 0}, & \eta_n \neq \eta_{n-1}. \end{cases} \tag{7}$$

Here h_n is necessary to memorize the current state of the voice part opposite of η_n. The minimization of the potential can be done with dynamic programming incrementally for each n. The estimation result is the same as the one with the Viterbi algorithm applied to the model when N_h is sufficiently large, and we confirmed that $N_h = 50$ is sufficient to provide a good approximation.

The algorithm was evaluated by applying it to several classical piano pieces. The used pieces were all pieces of Bach's two-voice inventions, the first two piano sonatas by Beethoven, Chopin's Etude Op. 10 Nos. 1–5, and the first three pieces in the first book of Debussy's Préludes. For comparison, we also evaluated algorithms based on lower-order HMMs. The zeroth-order model with transition and output probabilities $P(\eta)$ and $P(p|\eta)$ is almost equivalent to the keyboard splitting method, the first-order model with $P(\eta'|\eta)$ and $P(\delta p|\eta, \eta')$ and the second-order model are simple applications of HMMs whose latent variables are hand informations $\eta = L, R$.

The results are shown in Table 1. In total, the merged-output HMM yielded the lowest error rate, with which relatively accurate voice part separation can be done. On the other hand, there were less changes in results for the lower-order HMMs, showing that the effectiveness of the merged-output HMM. In Debussy's pieces, the error rates were relatively high since the pieces necessitate complex fingerings with wide movements of the hands. An example of the voice-part separation result is shown in Fig. 3.



(a) Passage in Bach's two-voice invention No. 1.

(b) Piano role representation of the voice-part separation result. Two voice parts are colored red and blue.

Figure 3. Example of a voice-part separation result.

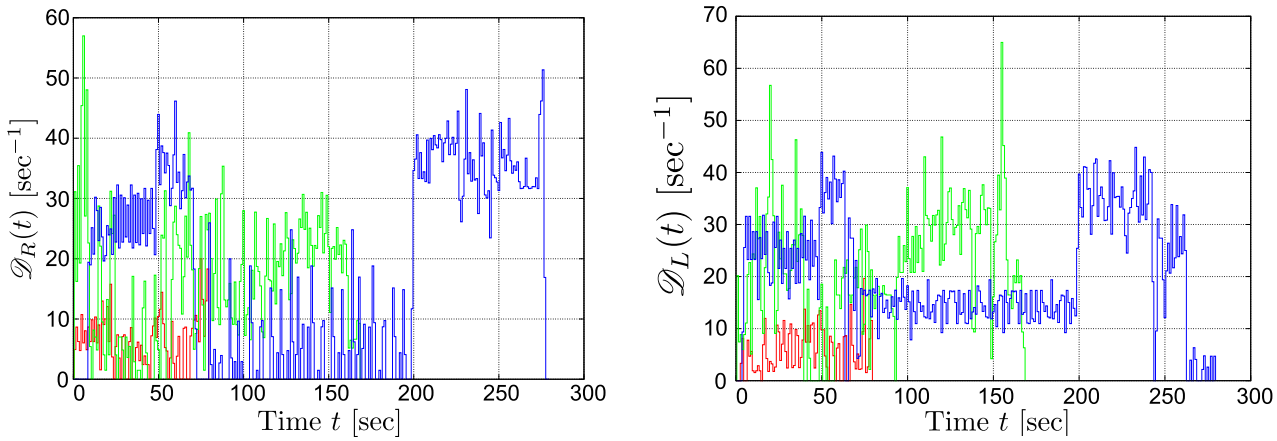
3.3 Quantitative measure of difficulty of performance

A measure of performance difficulty based on the naturalness of the fingerings can be obtained by the probabilistic fingering model. Although global structures in scores may influence the difficulty, we concentrate on the effect of local structures. It is supposed that the difficulty is additive with regard to performed notes and an increasing function of tempo. A quantity satisfying these conditions is the time rate of probabilistic cost. Let $\boldsymbol{p}(t)$ denote the sequence of notes in the time range of $[t - \Delta t/2, t + \Delta t/2]$, and $\boldsymbol{f}(t)$ be the corresponding fingerings, where Δt is a width of the time range to define the time rate. Then it is given as

$$\mathscr{D}(t) = -\ln P(\boldsymbol{p}(t), \boldsymbol{f}(t))/\Delta t. \tag{8}$$

Since the minimal time interval of successive notes are about a few 10 milli seconds and it is hard to imagine that difficulty is strongly influenced by notes that are separated more than 10 seconds, it is natural to set Δt within these extremes. The right-hand side is given by Eq. (1). It is possible to calculate $\mathscr{D}(t)$ for a score without indicated fingerings by replacing $\boldsymbol{f}(t)$ with the estimated fingerings $\hat{\boldsymbol{f}}(t)$ with the model in Sec. 2. In addition to the difficulty for both hands, that for each hand $\mathscr{D}_{L,R}(t)$ can also be defined similarly.

Fig. 4 shows some examples of $\mathscr{D}_{L,R}(t)$ calculated for several piano pieces. Here Δt was set to 1 sec. Although it is not easy to evaluate the quantity in a strict way, the results seems reasonable and reflects generic intuition of difficulty. The invention by Bach that can be played by beginners yields $\mathscr{D}_{L,R}$ that are less than about 10, the example of Beethoven's sonata which requires middle-level technicality has $\mathscr{D}_{L,R}$ around 20 to 30, and Chopin's Fantasie Impromptu which involves fast passages and difficult fingerings has $\mathscr{D}_{L,R}$ up to about 40. It is also worthy of noting that relatively difficult passages such as the fast chromatique passage of the right hand in the introduction of Beethoven's sonata and ornaments in the right hand of the



(a) Difficulty for right hand $\mathscr{D}_R$

(b) Difficulty for left hand $\mathscr{D}_L$

Figure 4. Examples of $\mathscr{D}_R$ and $\mathscr{D}_L$. The red (resp. green, blue) line is for Bach's two-voice invention No.=1, (resp. Introduction and exposition parts of the first movement of Beethoven's eighth piano sonata, Chopin's Fantasie Impromptu).

slow part of the Fantasie Impromptu are also captured in terms of $\mathscr{D}_R$.

4. CONCLUSIONS

In this paper, we considered a piano fingering model of both hands and its applications especially toward a piano reduction algorithm. First we reviewed a piano fingering model for one hand based on HMM, and then constructed a model for both hands based on merged-output HMM. Next we applied the model for constructing an algorithm for fingering decision and voice-part-separation algorithm and obtained a measure of performance difficulty. The algorithm for fingering decision yielded better results than the previously proposed one by a modification in details of the model. The results of voice-part separation is quite good and encouraging. The proposed measure of performance difficulty successfully captures the dependence on tempos and complexity of pitches and finger movements.

The next step to construct a piano reduction algorithm according to the formalization mentioned in the Introduction is to quantify the fidelity of the arranged score to the original score and to integrate it with the constraints of performance difficulty. The fidelity can be described with edit probability, similarly as in Ref. [5], and an arrangement model can be obtained by integrating the fingering model with the edit probability. We are currently working on these issues and the results will be reported elsewhere.

5. ACKNOWLEDGMENTS

This work is supported in part by Grant-in-Aid for Scientific Research from Japan Society for the Promotion of Science, No. 23240021, No. 26240025 (S.S. and N.O.), and No. 25880029 (E.N.).

6. REFERENCES

[1] S.-C. Chiu and M.-S. Chen, "A study on difficulty level recognition of piano sheet music," *Proc. AdMIRe*, pp. 10–12, 2012.

[2] S.-C. Chiu *et al.*, "Automatic system for the arrangement of piano reductions," *Proc. AdMIRe*, 2009.

[3] K. Fujita *et al.*, "A proposal for piano score generation that considers proficiency from multiple part (in Japanese)," *Tech. Rep. SIGMUS*, MUS-77, pp. 47–52, 2008.

[4] M. Hart and E. Tsai, "Finding optimal piano fingerings," *The UMAP Journal*, **21(1)**, pp. 167–177, 2000.

[5] G. Hori *et al.*, "Input-output HMM applied to automatic arrangement for guitars," *J. Information Processing*, **21(2)**, pp. 264–271, 2013.

[6] Z. Ghahramani and M. Jordan, "Factorial Hidden Markov Models," *Machine Learning*, **29**, pp. 245–273, 1997.

[7] A. Al Kasimi *et al.*, "A simple algorithm for automatic generation of polyphonic piano fingerings," *ISMIR*, pp. 355–356, 2007.

[8] F. Liszt, *Musikalische Werke*, Serie IV, Breitkopf & Härtel, 1922.

[9] J. Musafia, *The Art of Fingering in Piano Playing*, MCA Music, 1971.

[10] E. Nakamura *et al.*, "Merged-output hidden Markov model and its applications to score following and hand separation of polyphonic keyboard music (in Japanese)," *Tech. Rep. SIGMUS*, 2013-EC-27, **15**, 2013.

[11] E. Nakamura *et al.*, "Merged-output hidden Markov model for score following of MIDI performance with ornaments, desynchronized voices, repeats and skips," to appear in *Proc. ICMC*, 2014.

[12] C. Palmer, "Music performance," *Ann. Rev. Psychol.*, **48**, pp. 115–138, 1997.

[13] R. Parncutt *et al.*, "An ergonomic model of keyboard fingering for melodic fragments," *Music Perception*, **14(4)**, pp. 341–382, 1997.

[14] R. Parncutt *et al.*, "Interdependence of right and left hands in sight-read, written, and rehearsed fingerings of parallel melodic piano music," *Australian J. of Psychology*, **51(3)**, pp. 204–210, 1999.

[15] V. Sébastien *et al.*, "Score analyzer: Automatically determining scores difficulty level for instrumental e-learning," *Proc. ISMIR*, 2012.

[16] H. Sekiguchi and S. Eiho, "Generating and displaying the human piano performance," **40(6)**, pp. 167–177, 1999.

[17] Y. Takegawa *et al.*, "Design and implementation of a real-time fingering detection system for piano performance," *Proc. ICMC*, pp. 67–74, 2006.

[18] Y. Yonebayashi *et al.*, "Automatic determination of piano fingering based on hidden Markov model (in Japanese)," *Tech. Rep. SIGMUS*, 2006-05-13, pp. 7–12, 2006.

[19] Y. Yonebayashi *et al.*, "Automatic decision of piano fingering based on hidden Markov models," *IJCAI*, pp. 2915–2921, 2007.

[20] Y. Yonebayashi *et al.*, "Automatic piano fingering decision based on hidden Markov models with latent variables in consideration of natural hand motions (in Japanese)," *Tech. Rep. SIGMUS*, MUS-71-29, pp. 179–184, 2007.