BAYESIAN AUDIO ALIGNMENT BASED ON A UNIFIED GENERATIVE MODEL OF MUSIC COMPOSITION AND PERFORMANCE

Akira Maezawa[1,2] **Katsutoshi Itoyama**[2] **Kazuyoshi Yoshii**[2] **Hiroshi G. Okuno**[3]
[1]Yamaha Corporation [2]Kyoto University [3]Waseda University
{amaezaw1,itoyama,yoshii}@kuis.kyoto-u.ac.jp, okuno@aoni.waseda.jp

ABSTRACT

This paper presents a new probabilistic model that can align multiple performances of a particular piece of music. Conventionally, dynamic time warping (DTW) and left-to-right hidden Markov models (HMMs) have often been used for audio-to-audio alignment based on a shallow acoustic similarity between performances. Those methods, however, cannot distinguish latent musical structures common to all performances and temporal dynamics unique to each performance. To solve this problem, our model explicitly represents two state sequences: a top-level sequence that determines the common structure inherent in the music itself and a bottom-level sequence that determines the actual temporal fluctuation of each performance. These two sequences are fused into a hierarchical Bayesian HMM and can be learned at the same time from the given performances. Since the top-level sequence assigns the same state for note combinations that repeatedly appear within a piece of music, we can unveil the latent structure of the piece. Moreover, we can easily compare different performances of the same piece by analyzing the bottom-level sequences. Experimental evaluation showed that our method outperformed the conventional methods.

1. INTRODUCTION

Multiple audio alignment is one of the most important tasks in the field of music information retrieval (MIR). A piece of music played by different people produces different expressive performances, each embedding the unique interpretation of the player. To help a listener better understand the variety of interpretation or discover a performance that matches his/her taste, it is effective to clarify how multiple performances differ by using visualization or playback interfaces [1–3]. Given multiple musical audio signals that play a same piece of music from the beginning to the end, our goal is to find a temporal mapping among different signals while considering the underlying music score.

This paper presents a statistical method of offline multiple audio alignment based on a probabilistic generative model that can integrate various sources of uncertainties in music, such as spectral shapes, temporal fluctuations and structural deviations. Our model expresses how a *musical composition* gets *performed*, so it must model how they are generated.[1] Such a requirement leads to a conceptual model illustrated in Figure 1, described using a combination of two complementary models.

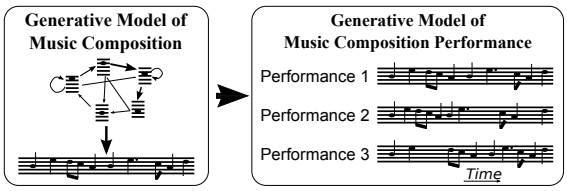

Figure 1. An overview of generative audio alignment.

To represent the generative process of a musical composition, we focus on the general fact that small fragments consisting of multiple musical notes form the basic building blocks of music and are organized into a larger work. For example, the sonata form is based on developing two contrasting fragments known as the "subject groups," and a song form essentially repeats the same melody. Our model is suitable for modeling the observation that basic melodic patterns are reused to form the sonata or the song.

To represent the generative process of each performance, we focus on temporal fluctuations from a common music composition. Since each performance plays the same musical composition, the small fragments should appear in the same order. On the other hand, each performance can be played by a different set of musical instruments with a unique tempo trajectory.

Since both generative processes are mutually dependent, we integrate a generative model of music composition with that of performance in a hierarchical Bayesian manner. In other words, we separate the characteristics of a given music audio signal into those originating from the underlying music score and those from the unique performance. Inspired by a typical preprocessing step in music structure segmentation [6,7], we represent a music composition as a sequence generated from a compact, ergodic Markov model ("latent composition"). Each music performance is represented as a left-to-right Markov chain that traverses the latent composition with the state durations unique to each performance.[2]

[1] A generative audio alignment model depends heavily on the model of both how the music is *composed* and how the composition is *performed*. This is unlike generative audio-to-score alignment [4,5], which does not need a music composition model because a music score is already given.

[2] Audio samples are available on the website of the first author.

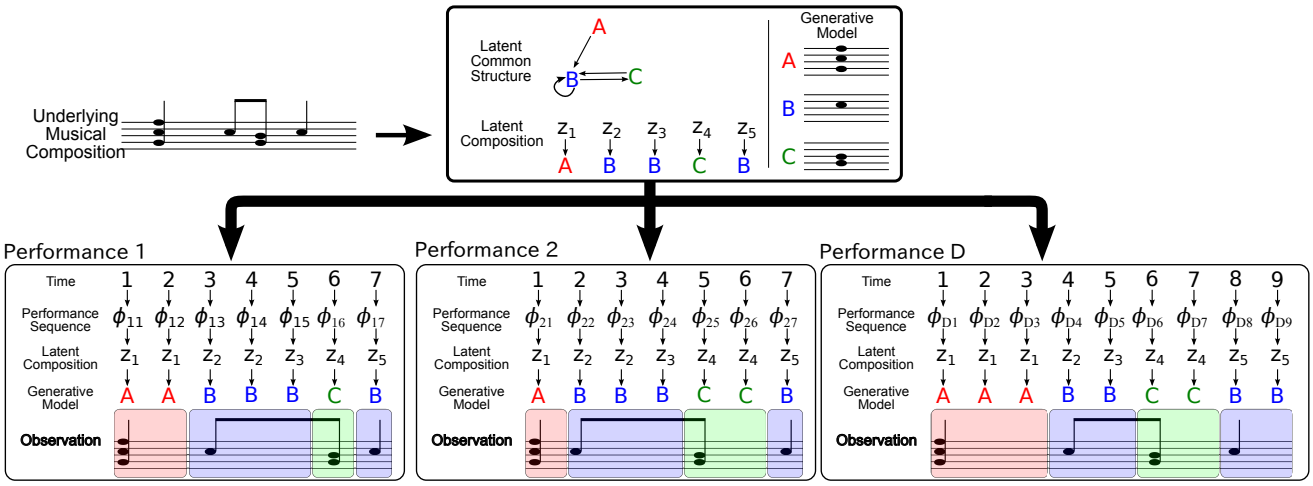

Figure 2. The concept of our method. Music composition is modeled as a sequence (composition sequence) from an ergodic Markov model, and each performance plays the composition sequence, traversing the composition sequence in the order it appears, but staying in each state with different duration.

2. RELATED WORK

Audio alignment is typically formulated as a problem of maximizing the similarity or minimizing the cost between a performance and another performance whose time-axis has been "stretched" by a time-dependent factor, using dynamic time warping (DTW) and its variants [8, 9] or other model of temporal dynamics [10]. To permit the use of a simple similarity measure, it is important to design robust acoustic features [11, 12].

Alternatively, tackling alignment by a probabilistic generative model has gathered attention, especially in the context of audio-to-music score alignment [4, 5]. In general, a probabilistic model is formulated to describe how each note in a music score translates to an audio signal. It is useful when one wishes to incorporate, in a unified framework, various sources of uncertainties present in music, such as inclusion of parts [13], mistakes [14], or timbral variations [15–17].

Previous studies in generative audio alignment [13, 18] ignores the organization present in musical composition, by assuming that a piece of music is generated from a left-to-right Markov chain, *i.e.*, a Markov chain whose state appears in the same order for all performances.

3. FORMULATION

We formulate a generative model of alignment that aligns D performances. We provide a conceptual overview, and then mathematically formalize the concept.

3.1 Conceptual Overview

We first extract short-time audio features from each of D performances. Let us denote the feature sequence for the dth performance at frame $t \in [1, T_d]$ as $\boldsymbol{x}_{d,t}$, where T_d is the total number of frames for the dth audio signal. Here, the kind of feature is arbitrary, and depends on the generative model of the short-time audio. Then, we model $\boldsymbol{x}_{d,t}$ as a set of D state sequences. Each state is associated with a unique generative process of short-time audio feature. In other words, each state represents a distinct audio feature, *e.g.*, distinct chord, f_0 and so on, depending on how the generative model of the feature is designed.

For audio alignment, the state sequence must abide by two rules. First, the order in which each state appears is the same for all D feature sequences. In other words, every performance is described by one sequence of distinct audio features, *i.e.*, the musical piece that the performances play in common. We call such a sequence the *latent composition*. Second, the duration that each performance resides in a given state in the latent composition can be unique to the performance. In other words, each performance traverses the latent composition with a unique "tempo curve." We call the sequence that each performance traverses over the latent composition sequence as the *performance sequence*.

The latent composition is a sequence of length N drawn from an ergodic Markov model, which we call the *latent common structure*. We describe the latent composition as $\boldsymbol{z}_n$, a sequence of length N and S states, where each state describes a distinct audio feature. In other words, we assume that the musical piece is described by at most N distinct audio events, using at most S distinct sounds. The latent common structure encodes the structure inherent to the music. The transition probabilities of each state sheds light on a "typical" performance, *e.g.*, melody line or harmonic progression. Therefore, the latent common structure provides a generative model of music composition.

The performance sequence provides a generative model of performance. Each audio signal is modeled as an emission from a N-state left-to-right Markov model, where the nth state refers to the generative model associated with the nth position in the latent composition. Specifically, let us denote the performance sequence for audio d as $\boldsymbol{\phi}_{d,t}$, which is a state sequence of length T_d and N states, such that state n refers to the nth element of the latent composition. Each performance sequence is constrained such that (1) it begins in state 1 and ends at state N, and (2) state n may traverse only to itself or state $n+1$. In other words, we

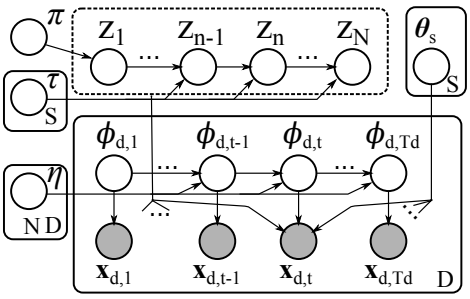

Figure 3. Graphical model of our method. Dotted box indicates that the arrow depends on all variables inside the dotted box. Hyperparameters are omitted.

constrain each performance sequence to traverse the latent composition in the same order but with a unique duration. Such a model conveys the idea that each performance can independently play a piece in any tempo trajectory.

3.1.1 An Example

Let us illustrate our method in Figure 2. In the example, $S = 3$ and $N = 5$, where state "A" corresponds to a combination of notes G, C and F, "B" corresponds to the note C, and so on; moreover, $\boldsymbol{z}_n$ encodes the state sequence "ABBCB," as to reflect the underlying common music composition that the performances play. Note that a single note may be expressed using more than one state in the latent composition, *e.g.*, both $\boldsymbol{z}_2$ and $\boldsymbol{z}_3$ describe the note "C." Next, each performance aligns to the latent composition, through the performance sequence. Each state of the performance sequence is associated to a position in the latent composition. For example, $\phi_{1,3}$ is associated to position 2 of $\boldsymbol{z}$, $\boldsymbol{z}_2$. Then, at each time, the observation is generated by emitting from the state in latent common structure referred by the current frame of the current audio. This is determined hierarchically by looking up the state n of the performance sequence of audio d at time t, and referring to the state s of the nth element of the latent composition. In the example, $\phi_{1,3}$ refers to state $n = 2$, so the generative model corresponding to $\boldsymbol{z}_{n=2}$, or "B," is referred.

3.2 Formulation of the Generative Model

Let us mathematically formalize the above concept using a probabilistic generative model, summarized as a graphical model shown in Fig. 3.

3.2.1 Latent Composition and Common Structure

The latent composition is described as $\boldsymbol{z}_{n=\{1\cdots N\}}$, a S-state state sequence of length N, generated from the latent common structure. We shall express the latent composition $\boldsymbol{z}_n$ using one-of-S representation; $\boldsymbol{z}_n$ is a S-dimensional binary variable where, when the state of $\boldsymbol{z}_n$ is s, $z_{n,s} = 1$ and all other elements are 0. Then, we model $\boldsymbol{z}$ as a sequence from the latent common structure, an ergodic Markov chain with initial state probability π and transition probability τ:

$$p(\boldsymbol{z}|\boldsymbol{\pi},\boldsymbol{\tau}) = \prod_{s=1}^{S} \pi_s^{z_{1,s}} \prod_{n=2,s'=1,s=1}^{N,S,S} \tau_{s,s'}^{z_{n-1,s'}z_{n,s}} \qquad (1)$$

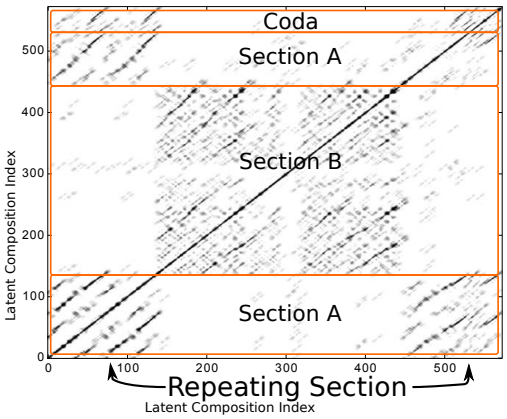

Figure 4. Structural annotation on Chopin Op. 41-2 and the similarity matrix computed from its latent composition.

Each state s is associated with an arbitrary set of parameters $\boldsymbol{\theta}_s$ that describes the generative process of the audio feature. We assume that $\boldsymbol{\tau}_s$ is generated from a conjugate Dirichlet distribution, *i.e.*, $\boldsymbol{\tau}_s \sim \mathrm{Dir}(\boldsymbol{\tau}_{0,s})$. The same goes for the initial state probability $\boldsymbol{\pi}$, *i.e.*, $\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\pi}_0)$. The hyperparameters $\boldsymbol{\tau}_{0,s}$ and $\boldsymbol{\pi}_0$ are set to a positive value less than 1, which induces sparsity of $\boldsymbol{\tau}$ and $\boldsymbol{\pi}$, and hence leads to a compact latent common structure.

The latent composition and structure implicitly convey the information about how the music is structured and what its building blocks are. Figure 4 shows a similarity matrix derived from the estimated latent composition of Op. 41-2 by F. Chopin[3] having the ternary form (*a.k.a.* ABA form). The first "A" section repeats a theme of form "DEDF" repeated twice. The second section is in a modulated key. Finally, the last section repeats the first theme, and ends with a short coda, borrowing from "F" motive from the first theme. Noting that the diagonal lines of a similarity matrix represent strong similarity, we may unveil such a trend by analyzing the matrix. The bottom-left diagonal lines in the first section, for example, shows that a theme repeats, and the top-left diagonal suggests that the first theme is repeated at the end. This suggests that the latent composition reflects the organization of music.

Notice that this kind of structure arises because we explicitly model the organization of music, conveyed through an ergodic Markov model; simply aligning multiple performances to a single left-to-right HMM [13, 18] is insufficient because it cannot revisit a previously visited state.

3.2.2 Performance Sequence

Recall that we require the performance sequence such that (1) it traverses in the order of latent composition, and (2) the duration that each performance stays in a particular state in the latent composition is conditionally independent given the latent composition. To satisfy these requirements, we model the performance sequence as a N-state left-to-right Markov chain of length T_d, $\phi_{d,t}$, where the first state of the chain is fixed to the beginning of the latent

[3] The similarity matrix $R_{i,j}$ was determined by removing self-transitions from $\boldsymbol{z}_n$ and assigning it to $\boldsymbol{z}'$, and setting $R_{i,j} = 1$ if $\boldsymbol{z}'_i = \boldsymbol{z}'_j$, and 0 otherwise. Next, we convolved R by a two-dimensional filter that emphasizes diagonal lines.

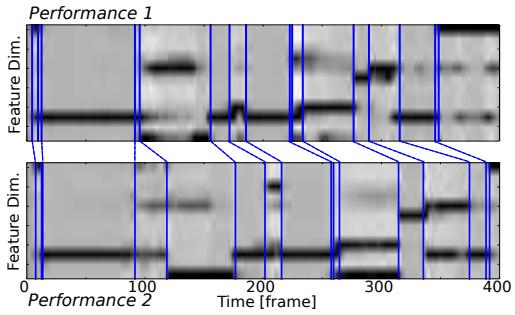

Figure 5. Feature sequences (chroma vector) of two performances, overlayed by points where the state of the latent composition changes.

composition and the last state to be the end. This assumes that there are no cuts or repeats unique to a performance. Let us define $\eta_{d,n}$ to be the probability for performance d to traverse from position n of the latent composition to $n+1$. Then, we model the performance sequence as follows:

$$p(\boldsymbol{\phi}_{d,t=\{1\cdots T_d\}}) = \delta(n,1)^{\phi_{d,1,n}}\delta(n,S)^{\phi_{d,T_d,n}} \times \prod_{t=1,n=1}^{T_d,N}\Big[\eta_{d,n}^{\phi_{d,t-1,n}\phi_{d,t,n+1}} \times (1-\eta_{d,n})^{\phi_{d,t-1,n}\phi_{d,t,n}}\Big] \quad (2)$$

where $\delta(x,y)$ indicates the Kronecker Delta, *i.e.*, its value is 1 when $x=y$ and 0 otherwise. We assume $\eta_{d,n}$ is drawn from a conjugate Beta distribution, *i.e.*, $\eta_{d,n} \sim \mathrm{Beta}(a_0, b_0)$. The ratio a_0/b_0 controls the likelihood of traversing to next states, and their magnitudes control the influence of the observation on the posterior distribution.

Figure 5 shows excerpts of the feature sequences obtained from two performances, and blue lines indicating the change of the state of the latent composition has changed. The figure suggests that the state changes with a notable change in the feature, such as when new notes are played. Since, by the definition of a left-to-right Markov model, the number of vertical lines is identical for all performances, we can align audio signals by mapping the occurrences of the ith vertical line for all performances, for each i.

3.2.3 Generating Audio Features

Based on the previous expositions, we can see that at time t of performance d, the audio feature is generated by choosing the state in the latent common structure that is referred at time t for performance d. This state is extracted by referring to the performance sequence to recover the position of the latent composition. Therefore, the observation likelihood is given as follows:

$$p(\boldsymbol{x}_{d,t}|\boldsymbol{z},\boldsymbol{\phi},\boldsymbol{\theta}) = \prod_{s,n} p(\boldsymbol{x}_{d,t}|\boldsymbol{\theta}_s)^{z_{n,s}\phi_{d,t,n}} \quad (3)$$

Here, $p(\boldsymbol{x}|\boldsymbol{\theta}_s)$ is the likelihood of observation feature $\boldsymbol{x}$ at state s of the latent common structure, and its parameter $\boldsymbol{\theta}_s$ is generated from a prior distribution $p(\boldsymbol{\theta}_s|\boldsymbol{\theta}_0)$.

For the sake of simplicity, we let $p(\boldsymbol{x}_{d,t}|\boldsymbol{\theta}_s)$ be a $\dim(\boldsymbol{x})$-dimensional Gaussian distribution with its parameters $\boldsymbol{\theta}_s$ generated from its conjugate distribution, the Gaussian-Gamma distribution. Specifically we let $\boldsymbol{\theta}_s = \{\boldsymbol{\mu}_s, \boldsymbol{\lambda}_s\}$, $\boldsymbol{\theta}_0 = \{\boldsymbol{m}_0, \nu_0, u_0, \boldsymbol{k}_0\}$, and let $\boldsymbol{x}_{d,t}|\boldsymbol{\mu}_s,\boldsymbol{\lambda}_s \sim \mathcal{N}(\boldsymbol{\mu}_s, \boldsymbol{\lambda}_s^{-1})$, with $p(\mu_{s,i},\lambda_{s,i}) \propto \lambda_s^{u_0-\frac{1}{2}} e^{-(\mu_{s,i}-m_{0,i})^2\lambda_{s,i}\nu_0 - k_{0,i}\lambda_{s,i}}$. One may incorporate a more elaborate model that better expresses the observation.

3.3 Inferring the Posterior Distribution

We derive the posterior distribution to the model described above. Since direct application of Bayes' rule to arrive at the posterior is difficult, we employ the variational Bayes method [19] and find an approximate posterior of form $q(\boldsymbol{\phi},\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\pi},\boldsymbol{\tau}) = \prod_d q(\boldsymbol{\phi}_{d,\cdot})q(\boldsymbol{z})q(\boldsymbol{\pi})\prod_{d,n} q(\eta_{d,n}) \prod_s q(\boldsymbol{\theta}_s)q(\boldsymbol{\tau}_s)$ that minimizes the Kullback-Leibler (KL) divergence to the true posterior distribution.

$q(\boldsymbol{\phi})$ and $q(\boldsymbol{z})$ can be updated in a manner analogous to a HMM. For $q(\boldsymbol{z})$, we perform the forward-backward algorithm, with the state emission probability $\boldsymbol{g}_n$ at position n of the latent composition and the transition probability $\boldsymbol{v}_s$ from state s given as follows:

$$\log g_{n,s} = \sum_{d,t}\langle\phi_{d,t,n}\rangle\langle\log p(\boldsymbol{x}_{d,t}|\boldsymbol{\theta}_s)\rangle \quad (4)$$

$$\log v_{s,s'} = \langle\log\tau_{s,s'}\rangle \quad (5)$$

Here, $\langle f(x)\rangle$ denotes the expectation of $f(x)$ w.r.t. q. Likewise, for $q(\boldsymbol{\phi}_{d,t})$, we perform the forward-backward algorithm, with the state emission probability $\boldsymbol{h}_{d,n}$ and transition probability $\boldsymbol{w}_{d,s}$ given as follows:

$$\log h_{d,t,n} = \sum_s \langle z_{n,s}\rangle\langle\log p(\boldsymbol{x}_{d,t}|\boldsymbol{\theta}_s)\rangle \quad (6)$$

$$\log w_{d,n,n'} = \begin{cases}\langle\log\eta_{d,n}\rangle & n = n' \\ \langle\log(1-\eta_{d,n})\rangle & n+1 = n'\end{cases} \quad (7)$$

We can update $\boldsymbol{\pi}$ as $q(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}_0 + \langle\boldsymbol{z}_1\rangle)$, $\boldsymbol{\eta}$ as $q(\eta_{d,n}) = \mathrm{Beta}(a_0 + \sum_t\langle\phi_{d,t-1,n}\phi_{d,t,n}\rangle, b_0 + \sum_t\langle\phi_{d,t-1,n-1}\phi_{d,t,n}\rangle)$, and $\boldsymbol{\tau}$ as $q(\boldsymbol{\tau}_s) = \mathrm{Dir}(\boldsymbol{\tau}_{0,s} + \sum_{n>1}^{N}\langle z_{n-1,s}\boldsymbol{z}_n\rangle)$.

Based on these parameters, the generative model of audio features can be updated. Some commonly-used statistics for state s include the count $\bar{N}_s$, the mean $\bar{\boldsymbol{\mu}}_s$ and the variance $\bar{\boldsymbol{\Sigma}}_s$, which are given as follows:

$$\bar{N}_s = \sum_{d,n,t}\langle z_{n,s}\rangle\langle\phi_{d,t,n}\rangle \quad (8)$$

$$\bar{\boldsymbol{\mu}}_s = \frac{1}{\bar{N}_s}\sum_{d,n,t}\langle z_{n,s}\rangle\langle\phi_{d,t,n}\rangle\boldsymbol{x}_{d,t} \quad (9)$$

$$\bar{\boldsymbol{\Sigma}}_s = \frac{1}{\bar{N}_s}\sum_{d,n,t}\langle z_{n,s}\rangle\langle\phi_{d,t,n}\rangle(\boldsymbol{x}_{d,t}-\bar{\boldsymbol{\mu}}_s)^2 \quad (10)$$

For example, the Gaussian/Gaussian-Gamma model described earlier can be updated as follows:

$$q(\boldsymbol{\mu}_s,\boldsymbol{\lambda}_s) = \mathcal{NG}\Big(\nu_0+\bar{N}_s, \frac{\nu_0\boldsymbol{m}_0+\bar{N}_s\bar{\boldsymbol{\mu}}_s}{\nu_0+\bar{N}_s}, u_0+\frac{\bar{N}_s}{2}, \boldsymbol{k}_0+\frac{1}{2}\Big(\bar{N}_s\bar{\boldsymbol{\Sigma}}_s+\frac{\nu_0\bar{N}_s}{\nu_0+\bar{N}_s}(\bar{\boldsymbol{\mu}}_s-\boldsymbol{m}_0)^2\Big)\Big) \quad (11)$$

Hyperparameters may be set manually, or optimized by minimizing the KL divergence from q to the posterior.

3.4 Semi-Markov Performance Sequence

The model presented previously implicitly assumes that the state duration of the performance sequence follows the

geometric distribution. In such a model, it is noted, especially in the context of audio-to-score alignment [4], that further improvement is possible by incorporating a more explicit duration probability using an extension of the HMM known as the hidden semi-Markov models [5, 20].

In this paper, we assume that every performance plays a *particular position* in the music composition with more-or-less the same tempo. Hence, we incorporate an explicit duration probability to the performance sequence, such that the duration of each state is concentrated about some average state duration common to each performance. To this end, we assume that for each state n of the performance sequence, the state duration l follows a Gaussian distribution concentrated about a common mean:

$$p(l|\gamma_n, c) = \mathcal{N}(\gamma_n, c\gamma_n^2) \tag{12}$$

We chose the Gaussian distribution due to convenience of inference. By setting c appropriately, we can provide a trade-off between the tendency for every piece to play in a same tempo sequence, and variation of tempo among different performances.

To incorporate such a duration probability in the performance sequence model, we augment the state space of the left-to-right Markov model of the performance sequence by a "count-down" variable l that indicates the number of frames remaining in the current state. Then, we assume that the maximum duration of each state is L, and represent each state of the performance $\boldsymbol{\phi}_{d,t}$ as a tuple $(n, l) \in [1 \cdots N] \times [1 \cdots L]$, *i.e.*, $\phi_{d,t,n,l}$. In this model, state $(n, 1)$ transitions to $(n+1, l)$ with probability $p(l|\mu_{n+1}, c)$, and state (n, l) for $l > 1$ transitions to $(n, l-1)$ with probability one. Finally, we constrain the terminal state to be $(N, 1)$. Note that η is no longer used because state duration is now described explicitly. The parameter γ_n can be optimized by maximum likelihood estimation of the second kind, to yield the following:

$$\gamma_n = \frac{\sum_{d,t,l} l\langle\phi_{d,t-1,n-1,1}\phi_{d,t,n,l}\rangle}{\sum_{d,t,l}\langle\phi_{d,t-1,n-1,1}\phi_{d,t,n,l}\rangle} \tag{13}$$

c may be optimized in a similar manner, but we found that the method performs better when c is fixed to a constant.

4. EVALUATION

We conducted two experiments to assess our method. First, we tested the effectiveness of our method against existing methods that ignore the organization of music [13, 18]. Second, we tested the robustness of our method to the length of the latent composition, which we need to fix in advance.

4.1 Experimental Conditions

We prepared two to five recordings to nine pieces of Chopin's *Mazurka* (Op. 6-4, 17-4, 24-2, 30-2, 33-2, 41-2, 63-3, 67-1, 68-3), totaling in 38 audio recordings. For each of the nine pieces, we evaluated the alignment using (1) DTW using path constraints in [21] that minimizes the net squared distance (denoted "DTW"), (2) left-to-right HMM to model musical audio as done in existing methods [13, 18] (denoted "LRHMM"), (3) proposed method (denoted "Proposed"), and (4) proposed method with semi-Markov performance sequence (denoted "Proposed (HSMM)"). For the feature sequence $\boldsymbol{x}_{d,t}$, we employed the chroma vector [11] and half-wave rectified difference of the chroma (Δ chroma), evaluated using a frame length of 8192 samples and a 20% overlap with a sampling frequency of 44.1kHz.

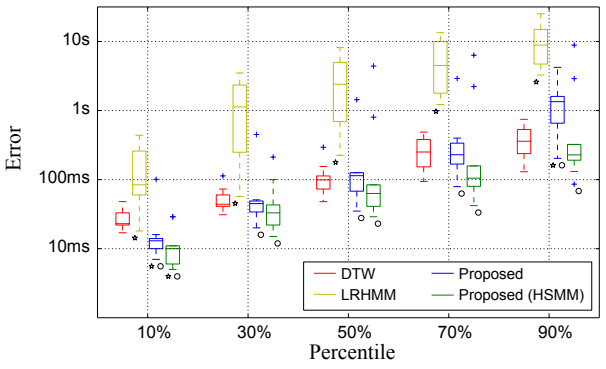

Figure 6. Percentile of absolute alignment error. Asterisks indicate statistically significant difference over DTW (p=0.05) and circles indicate statistically significant difference over LRHMM (p=0.05), using Kruskal-Wallis H-test.

For the proposed method, the hyperparameters related to the latent common structure were set to $\boldsymbol{\pi}_0 = 0.1$ and $\tau_{0,s,s'} = 0.9 + 10\delta(s, s')$; these parameters encourages sparsity of the initial state probability and the state transitions, while encouraging self-transitions. The parameters related to the observation were set to $u_0 = \boldsymbol{k}_0 = 1$, $\nu_0 = 0.1$ and $\boldsymbol{m}_0 = 0$; such a set of parameters encourages a sparse variance, and assumes that the mean is highly dispersed. Moreover, we used $S = 100$ and $N = 0.3 \min_d T_d$. For the semi-Markov performance sequence model, we set $c = 0.1$. This corresponds to having a standard deviation of $\gamma_n\sqrt{0.1}$, or allowing the notes to deviate by a standard deviation of about 30%.

4.2 Experimental Results

We present below the evaluation of the alignment accuracy and the robustness to the length of the latent composition. On a workstation with Intel Xeon CPU (3.2GHz), our method takes about 3 minutes to process a minute of single musical audio.

4.2.1 Alignment Accuracy

We compared the aligned data to that given by reverse conducting data of the Mazurka Project [1]. Figure 6 shows the absolute error percentile. The figure shows that our method ("Proposed") performs significantly better than the existing method based on a LRHMM. This suggests that, for a generative model approach to alignment, not only is model of performance *difference* critical but also that of the *common* music that the performances play. We also note an improved performance of the semi-Markov model performance sequence ("Proposed (HSMM)") over the Markovian model ("Proposed").

Note that when using the same features and squared-error model, the semi-Markovian model performs better

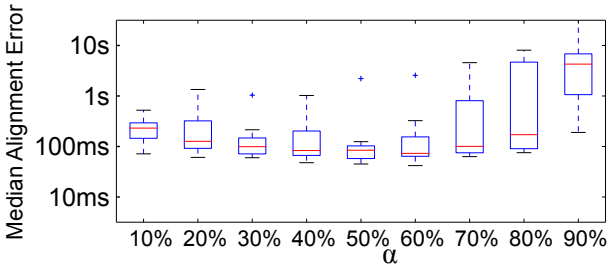

Figure 7. Median alignment error against α.

than DTW. This result suggests that with appropriate structural and temporal models, a generative model approach is a viable alternative to audio alignment. The performance gain from Markov to semi-Markov model illuminates the forte of the generative model approach: temporal, spectral and structural constraints are mixed seamlessly to attain a trade-off among the trichotomy.

We note that our model is weak to compositional deviations, such as added ornaments and repeats because we assume every performance plays an identical composition. We observed that our method deals with an added note as a noise or a note that gets played very shortly by most of the audio signals, but neither captures the nature of added notes as structural deviations. Moreover, our method sometimes gets "trapped" in local optima, most likely due to the strong mutual dependency between the latent variables.

4.2.2 Robustness to the Length of the Latent Composition

Since our method requires the user to set the length of latent composition N, we evaluated the quality of alignment as N is varied. To evaluate the performance of our method with different values of N, we evaluated the alignment of the proposed method when N is set to $N = \alpha|T_{d=1}|$, with α ranging from $\alpha = 0.1$ to $\alpha = 0.9$ with an increment of 0.1. Figure 7 shows the median alignment error. We find that when α is too small, when there is an insufficient number of states to describe a composition, the error increases. The error also increases when α is too large, since the maximum total allowed deviation decreases (*i.e.*, to about $(1-\alpha)T_{d=1}$). However, outside such extremities, the performance is relatively stable for moderate values of α around 0.5. This suggests that our method is relatively insensitive to a reasonable choice of N.

5. CONCLUSION

This paper presented an audio alignment method based on a probabilistic generative model. Based on the insight that a generative model of musical audio alignment should represent both the underlying musical composition and how it is performed by each audio signal, we formulated a unified generative model of musical composition and performance. The proposed generative model contributed to a significantly better alignment performance than existing methods. We believe that our contribution brings generative alignment on par with DTW-based alignment, opening door to alignment problem settings that require integration of various sources of uncertainties.

Future study includes incorporating better models of composition, performance and observation in our unified framework. In addition, inference over highly coupled hierarchical discrete state models is another future work.

Acknowledgment: This study was supported in part by JSPS KAKENHI 24220006 and 26700020.

6. REFERENCES

[1] C. S. Sapp. Comparative analysis of multiple musical performances. In *ISMIR*, pages 2–5, 2007.

[2] S. Miki, T. Baba, and H. Katayose. PEVI: Interface for retrieving and analyzing expressive musical performances with scape plots. In *SMC*, pages 748–753, 2013.

[3] C. Fremerey, F. Kurth, M. Müller, and M. Clausen. A demonstration of the SyncPlayer system. In *ISMIR*, pages 131–132, 2007.

[4] C. Raphael. A hybrid graphical model for aligning polyphonic audio with musical scores. In *ISMIR*, pages 387–394, 2004.

[5] A. Cont. A coupled duration-focused architecture for real-time music-to-score alignment. *IEEE PAMI*, 32(6):974–987, 2010.

[6] J. Paulus, M. Muller, and A. Klapuri. State of the art report: Audio-based music structure analysis. In *ISMIR*, pages 625–636, Aug. 2010.

[7] S. A. Abdallah et al. Theory and evaluation of a Bayesian music structure extractor. In *ISMIR*, pages 420–425, 2005.

[8] R. B. Dannenberg and N. Hu. Polyphonic audio matching for score following and intelligent audio editors. In *ICMC*, September 2003.

[9] M. Grachten et al. Automatic alignment of music performances with structural differences. In *ISMIR*, pages 607–612, 2013.

[10] N. Montecchio and A. Cont. A unified approach to real time audio-to-score and audio-to-audio alignment using sequential Monte-Carlo inference techniques. In *ICASSP*, pages 193–196, 2011.

[11] T. Fujishima. Realtime chord recognition of musical sound: A system using Common Lisp Music. In *ICMC*, pages 464–467, 1999.

[12] S. Ewert, M. Müller, and P. Grosche. High resolution audio synchronization using chroma onset features. In *ICASSP*, pages 1869–1872, 2009.

[13] A. Maezawa and H. G. Okuno. Audio part mixture alignment based on hierarchical nonparametric Bayesian model of musical audio sequence collection. In *ICASSP*, pages 5232–5236, 2014.

[14] T. Nakamura, E. Nakamura, and S. Sagayama. Acoustic score following to musical performance with errors and arbitrary repeats and skips for automatic accompaniment. In *SMC*, pages 200–304, 2013.

[15] A. Maezawa et al. Polyphonic audio-to-score alignment based on Bayesian latent harmonic allocation hidden Markov model. In *ICASSP*, pages 185–188, 2011.

[16] T. Otsuka et al. Incremental Bayesian audio-to-score alignment with flexible harmonic structure models. In *ISMIR*, pages 525–530, 2011.

[17] C. Joder, S. Essid, and G. Richard. Learning optimal features for polyphonic audio-to-score alignment. *IEEE TASLP*, 21(10):2118–2128, 2013.

[18] R. Miotto, N. Montecchio, and N. Orio. Statistical music modeling aimed at identification and alignment. In *AdMiRE*, pages 187–212, 2010.

[19] M. J. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, University College London, 2003.

[20] S. Yu and H. Kobayashi. An efficient forward-backward algorithm for an explicit-duration hidden Markov model. *IEEE SPL*, 10(1):11–14, Jan 2003.

[21] N. Hu, R. B. Dannenberg, and G. Tzanetakis. Polyphonic audio matching and alignment for music retrieval. In *WASPAA*, pages 185–188, 2003.