UNIFIED INTER- AND INTRA-RECORDING DURATION MODEL FOR MULTIPLE MUSIC AUDIO ALIGNMENT

Akira Maezawa[1] *Katsutoshi Itoyama*[2] *Kazuyoshi Yoshii*[2] *Hiroshi G. Okuno*[3]

1 Yamaha Corporation, Japan
2 Graduate School of Informatics, Kyoto University, Japan
3 Graduate School of Creative Science and Engineering, Waseda University, Japan

ABSTRACT

This paper presents a probabilistic audio-to-audio alignment method that focuses on the relationship among the note durations of different performances of a piece of music. A key issue in probabilistic audio alignment methods is in expressing how interrelated are the durations of notes in the underlying piece of music. Existing studies focus either on the duration of adjacent notes *within* a recording (intra-recording duration model), or the duration of a given note *across* different recordings (inter-recording duration model). This paper unifies these approaches through a simple modification to them. Furthermore, the paper extends the unified model, allowing the dynamics of the note duration to change sporadically. Experimental evaluation demonstrated that the proposed models decrease the alignment error.

***Index Terms*—** audio alignment, music information retrieval, hierarchical Bayesian model

1. INTRODUCTION

Multiple music audio-to-audio alignment is a task that locates where multiple audio renditions of a given piece of music are playing the same position in the piece. It is an important problem in music information retrieval. For example, audio-to-audio alignment [1–4] (or its score-informed cousin, audio-to-score alignment [5–11]) is useful for comparing different audio renditions of a given piece of music [12–14], since different audio renditions are played with different tempo trajectories. Furthermore, it is also potentially useful in wider applications if one could align audio signals that play significantly different parts of a same piece of music. For example, by aligning an audio track of a violin concerto and an audio track of the violin solo, one may apply separation-by-humming [15] to create a karaoke track of the concerto.

When aligning audio signals that contain highly varying signals, *e.g.*, solo violin track versus full orchestra track, probabilistic formulation to audio alignment [16] is preferred over more conventional approach based on path optimization [17]. Probabilistic formulation is preferable because it allows us to express, in a principled manner, the uncertainties underlying the spectral time-slices and its temporal evolution.

In probabilistic audio alignment, the model of the duration of each note (or note combination) of the underlying piece of music plays a critical role, in addition to the model of the common underlying piece of music. Unlike the path optimization approach, which involves matching of local audio features, probabilistic audio alignment typically needs to also infer the underlying generative model of audio features, complicating the problem. Thus, while the path optimization approach is robust without a duration model, probabilistic models merit by introducing elaborate note duration models. Such a model has been tackled by exploting two complementary aspects of note durations.

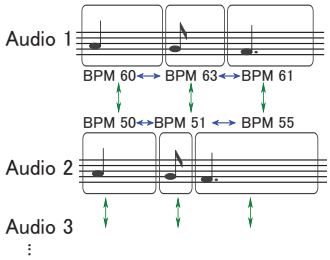

Figure 1: The concept behind the proposed duration model. We combine intra-recording constraints (blue arrows) and intra-recording constraints (green arrows).

First, it is possible to model the duration of adjacent notes *within* a piece of music, based on the insight that tempo tends to evolve smoothly over time [18]. In this approach, widely employed in audio-to-score alignment [6–11], one assumes an underlying smooth "tempo" curve. It is then combined with the information about the musical duration of each note (*e.g.* a sixteenth note versus a quarter note) to arrive at the (temporal) duration. In other words, it considers a dependency structure as illustrated by the blue arrows in Fig. 1. We call this the *intra-recording* duration model. This approach is effective when the tempo is more-or-less stationary, but has difficulties when there are abrupt changes of tempo.

Second possibility is to model the duration of a note *across* different recordings, based on the idea that a musically acceptable duration for a given note tends to lie within a narrow confine [19]. In this approach, one assumes that every recording plays a noisy rendition of an underlying tempo curve. In other words, it considers a dependency structure as illustrated by the green arrows in Fig. 1. We call this the *inter-recording* duration model. This approach works when different recordings play in similar tempi, but fails when the tempi vary significantly with different recordings.

To use the best of these complementary approaches, this paper first introduces a duration model that unifies intra- and inter-recording models. Then, to allow for sporadic changes in the tempo dynamics, we extend the unified duration model, as to allow the dynamics of the tempo to change sporadically.

This study was partially supported by JSPS KAKENHI 26700020, 24220006, 24700168.

978-1-4799-7450-4/15/$31.00 ©2015 IEEE

2. EXISTING MODELS

We shall briefly review the inter- and intra-recording duration models, as they lay the foundation for the proposed model. Let us assume that we are given I audio recordings of a same piece of music. We furthermore assume that we can segment the underlying music representation into N segments, where a given segment index of every audio refers to the same place in the underlying piece. We shall denote the duration of the nth segment of the ith recording as l_{ni}, and $\boldsymbol{l}_n$ as $\{l_{ni}\}_{i=1}^I$. The goal of alignment is to find $\boldsymbol{l}$ such that the nth segment of each signal represents the same position in the underlying piece. The segmentation may be partly given (*e.g.*, using a music score of the underlying piece, as in audio-to-score alignment), or treated as random variables and inferred in tandem with the segment durations.

In the intra-recording duration model for audio alignment [20], more popularly employed in audio-to-score alignment literatures [6–11], the tempo curve is assumed to be smooth. To reflect this assumption, we augment the duration model with an additional "tempo" variable, μ_{ni}, which indicates the *beat* duration of the nth segment for the ith recording. We shall denote $\boldsymbol{\mu}_n = \{\mu_{ni}\}_{i=1}^I$. Then, assuming $\mu_{n-1,i}$ is close to μ_{ni}, the duration $\{\boldsymbol{l}_n\}$ is expressed as a linear dynamical system (LDS) of the following form:

$$\boldsymbol{l}_n = a_n \boldsymbol{\mu}_n + \boldsymbol{\epsilon}_n^{(l)} \tag{1}$$

$$\boldsymbol{\mu}_n = \boldsymbol{\mu}_{n-1} + \boldsymbol{\epsilon}_n^{(\mu)}. \tag{2}$$

Here, a_n corresponds to the note duration of the nth segment (*e.g.*, a sixteenth note), and $\epsilon_n^{(l,\mu)}$ are the innovation of the duration and tempo, respectively. ϵ's are typically treated as zero-mean Gaussian random variables with small variances, due to its mathematical convenience and the intuition that that the change of tempo or duration is small. Intra-recording duration model is effective in many parts in a piece of music because the tempo tends to remain steady. On the other hand, the duration model does not work when the tempo wavers significantly.

As another strategy for multiple ($I \geq 2$) audio alignment, it is possible to focus on the property of $\{\boldsymbol{l}_n\}_{n=1}^N$ *across* different recordings (inter-recording duration model) [19]. This is based on the intuition that the range of musically acceptable tempo is relatively narrow, meaning that the duration of a particular note in a piece of music is relatively confined. Thus, one may reasonably assume that l_{ni} shares the same mean for different i's:

$$\boldsymbol{l}_n = m_n + \boldsymbol{\epsilon}_n^{(l)}. \tag{3}$$

The deviation from the average tempo curve, $\epsilon_{ni}^{(l)}$, is a zero-mean Gaussian random variable, whose standard deviation scales with the expected value of m_n. This kind of model induces coupling among l_{ni}'s across different i, encouraging every recording to take on a similar range of tempo curve. Since this method, unlike the intra-recording model, is free of assumption regarding the properties of tempo curves, the method is robust to wavering of tempo, as long as every recording wavers the tempo in a consistent manner. On the other hand, since it does not exploit a widely-applicable assumption of tempo smoothness, it may produce poorer alignment during segments with a smooth tempo curve.

3. THE PROPOSED METHOD

This paper presents two extensions to the inter- and intra-recording duration models. First, we unify the two models, allowing the duration model to simultaneously exploit properties within a recording and across recordings. Second, we extend the intra-recording and the unified duration model, allowing the tempo curve to switch abruptly in sporadic locations in the piece of music.

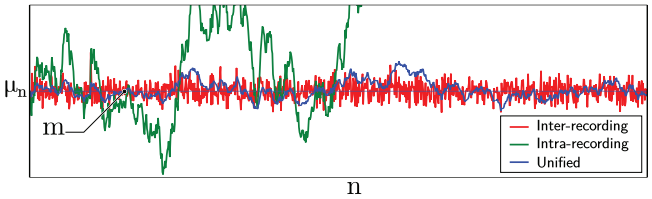

Figure 2: Realizations of the inter-recording, intra-recording and the unified duration models. Inter-recording (red) is close to the mean but is jumpy. Intra-recording (green) is smooth but may waver far away from the mean. The unified model (blue) is both smooth and close to the mean.

3.1. Unified duration model

Since inter- and intra-recording duration models focus on complementary aspects of the temporal progression of music, we expect that combining them would allow the two models to compensate for deficits of each other.

Thus, let us unify the inter- and intra-recording models. First, the inter-recording model is rewritten as follows:

$$\boldsymbol{l}_n = a_n \boldsymbol{\mu}_n + \boldsymbol{\epsilon}_n^{(l)} \tag{4}$$

$$\boldsymbol{\mu}_n = m + \boldsymbol{\epsilon}_n^{(\mu)}. \tag{5}$$

Next, let us take the weighted combination of Eq. 2 and Eq. 5, with the weight of inter-recording model specified as $\alpha \in [0, 1]$. Then, inter- and intra-recording duration models may be unified as a LDS of the following form:

$$\boldsymbol{l}_n = a_n \boldsymbol{\mu}_n + \boldsymbol{\epsilon}_n^{(l)} \tag{6}$$

$$\boldsymbol{\mu}_n = (1-\alpha)\boldsymbol{\mu}_{n-1} + \alpha m + \boldsymbol{\epsilon}_n^{(\mu)}. \tag{7}$$

We call this the *unified* duration model. In the unified duration model, the α mixes the balance between inter- and intra-recording models. Notice that we recover Eq. 2 by setting $\alpha = 0$ and Eq. 5 by setting $\alpha = 1$. We compare realizations of the inter-recording, the intra-recording and the unified model in Fig. 2.

We assume that the innovation $\boldsymbol{\epsilon}_{ni}^{(l)}$ is generated from a zero-mean Gaussian with inverse covariance (precision) λ_0. λ_0 controls how much deviation of the segment durations from the unified duration model is allowed. For example, agogic accent can be explained by a large $\epsilon_{ni}^{(l)}$.

We furthermore assume that $\boldsymbol{\epsilon}_n^{(\mu)}$ is generated from an I-dimensional, zero-mean Gaussian random variable with precision matrix $\boldsymbol{\Lambda}_n$. It governs how much the tempo may waver. If there are non-zero off-diagonal elements, it conveys the correlation in the increments of $\boldsymbol{\mu}$. We will present a more elaborate form of $\boldsymbol{\Lambda}_m$.

It is illuminating to analyze the stationary distribution of μ. For the sake of analysis, let us assume for the moment that $\boldsymbol{\epsilon}_n^{(\mu)}$ is a zero-mean Gaussian random variable with a spherical covariance, $\boldsymbol{\Lambda}_n^{-1} = \sigma^2 \boldsymbol{I}$. Then, the expectation of $\boldsymbol{\mu}$ is m, and the covariance is $\frac{\sigma^2}{\alpha(2-\alpha)}\boldsymbol{I}$. The finite variance suggests that μ_n tends to revert back to m for $\alpha > 0$. This property can be seen clearly by re-writing

Eq. 7 as follows:

$$\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n-1} = -\alpha(\boldsymbol{\mu}_{n-1} - m) + \boldsymbol{\epsilon}_n^{(\mu)}. \quad (8)$$

We can see that the first-order difference is directed towards m, and is proportional to how far $\boldsymbol{\mu}_{n-1}$ is from m. Thus, the farther away $\boldsymbol{\mu}$ is from m, the stronger is the restoring force back to m.

Based on the stationary distribution, one might set m to be the expected beat duration (as one might do in a inter-recording model); σ to be the expected deviation in tempo between adjacent beats (as one might do in a intra-recording model); and set α to reflect the expected *variation of tempo* among different performances.

3.1.1. Bayesian extension

The joint intra/inter-recording duration model can be easily extended for Bayesian analysis, by introducing appropriate prior distributions for parameters needs to be treated as random variables. Since a_n is an unknown quantity that depends on the nature of the inferred nth segment, we treat it as a zero-mean Gaussian random variable with a large variance ι^{-1}. In this paper, we fix m, α and λ_0. By fixing m, a_n infers the average duration of the nth segment, with μ_{ni} expressing the multiplicative offset from a_n.

3.2. Switching-state intra-recording/unified duration model

The intra-recording model underlying the unified duration model works when the tempo remains smooth, but the degree of smoothness may change abruptly during the piece. Such abrupt changes occur, for example, in structural boundaries [21].

To allow for such sporadic changes, we, inspired by tempo trajectory models [22], consider expressing $\boldsymbol{\mu}$ as a *switching-state* LDS (SSLDS). Let us assume that $\boldsymbol{\epsilon}_n$ is a Gaussian noise with zero mean and a precision matrix $\boldsymbol{\Lambda}_n$ chosen from one of M "patterns" of precision matrices, $\{\hat{\boldsymbol{\Lambda}}_m\}_{m=1}^M$. The pattern should remain stationary unless there is truly a change in the underlying dynamics. Thus, the sequence of the choice of the precision matrix $\boldsymbol{u} = \{u_n\}_{n=1}^N$ is expressed as a M-state Markov chain, with the state transition pdf $\{\boldsymbol{\xi}_m\}_{m=1}^M$:

$$\boldsymbol{\Lambda}_n = \tilde{\boldsymbol{\Lambda}}_{u_n} \quad (9)$$

$$u_n \sim \text{Discrete}(\boldsymbol{\xi}_{u_{n-1}}). \quad (10)$$

Preliminary analysis suggests that the state of $\boldsymbol{u}$ changes at structural boundaries of musical interpretation. To illustrate, we present the *maximum a posteriori* (MAP) estimate of $\boldsymbol{u}$ and $\hat{\boldsymbol{\Lambda}}$ for a short violin phrase shown in Fig. 3, played with 24 different tempo trajectories. Specifically, the dth recording is played by permuting over three binary decisions over the phrasing (dotted lines) and choosing an overall tempo (variable s in the figure). The MAP estimate of $\boldsymbol{u}$ is shown as color labels over the music score, and $\hat{\boldsymbol{\Lambda}}$ is shown in the bottom of the figure. It shows that performances with similar overall tempo are correlated, as shown by the block-like covariance structure. The red-colored label is a "default" covariance, and the remaining indicate spurious changes in the degree of smoothness. Note that the spurious changes (*i.e.*, non-red labels) often occur at phrase boundaries (*i.e.*, ends of arrows in Fig. 3).

This model naturally extends intra-recording duration models [6, 20], to deal with a sporadically varying degree of smoothness. Our use of the SSLDS is different from those conventionally used in beat tracking [23] or score following [24]. Namely, whereas these studies use the switching-state dynamics to express the *observation noise* (and a fixed tempo trajectory dynamics), our method usees the switching-state dynamics to express the *tempo trajectory*. Our method allows the underlying tempo to exhibit different dynamics, similar in spirit to expressive tempo modeling [22].

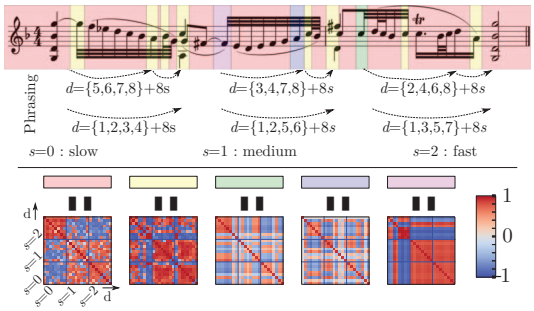

Figure 3: A short phrase played with 24 different interpretations, and the estimated $\boldsymbol{u}$ and $\hat{\boldsymbol{\Lambda}}$. Dotted arrows indicate the phrasing.

3.2.1. Bayesian extension

For a Bayesian extension, it suffices to introduce prior distributions over the state transition and the covariance matrices. Here, we assume that $\boldsymbol{\xi}_m$ is generated from a conjugate Dirichlet distribution, *i.e.*, $\boldsymbol{\xi}_m \sim \text{Dir}(\xi_0)$. We furthermore assume that $\hat{\boldsymbol{\Lambda}}_m$ is generated from a conjugate Wishart distribution, *i.e.*, $\hat{\boldsymbol{\Lambda}}_m \sim \text{W}(n_0, \boldsymbol{W}_0)$.

3.3. Applying the duration model for multiple audio alignment

The proposed duration model builds on top of a probabilistic multiple audio alignment method. Such an alignment method that the model builds upon should satisfy these requirements:

1. The method should take as inputs music audio signals (*i.e.*, without needing a symbolic music score)
2. The method should segment the input audio signals into chunks with similar-sounding segments
3. The method should be able to associate a duration pdf to each of the chunks.

To this end, we use [19], which satisfies these needs. Hereon we shall call this the *baseline* alignment method. In the baseline method, the input audio is segmented into similar-sounding fragments, allowing us to model the duration of similar-sounding segment of sound. Given I recordings, the method segments them into N segments, each of which is less than L frames. Here, nth segment for all recordings refer to the same position in the underlying piece of music. The segmentation is stored in variable z_{it}, which is one-of-$N \times L$ variable, where $z_{itnl} = 1$ means that audio i at time t is in segment n, with l frames remaining in that segment.

Note that, though not the focus of this paper, the proposed duration models may be applied to probabilistic audio-to-score alignment methods as well [6–9]. In this case, a_n should be specified from the symbolic score, and m_n possibly generated by analyzing the tempo markings of the symbolic score.

4. INFERENCE

We estimate the alignment by inferring the posterior distribution, using it to find the MAP estimate of z_{itnl}. Then we determine the frames that play the nth segment for each signal i .

Variational Bayesian method is used to approximate the posterior distribution [25]. For the unified duration model, we use structured mean-field approximation, by approximating the true posterior distribution with an approximate distribution in which $\{\boldsymbol{\mu}_n\}_{n=1}^N$, $\{\boldsymbol{a}_n\}_{n=1}^N$, and variables related to the baseline alignment method (such as z and other variables not introduced in this paper) are mutually independent. Then, we minimize the KL divergence from the approximate posterior to the true posterior, by iteratively minimizing each factor. For the switching-state unified duration model, we use the same structured variational inference, and furthermore assume mutual independence of $\{u_n\}_{n=1}^N$, $\{\boldsymbol{\xi}_m\}_{m=1}^M$, and $\{\hat{\boldsymbol{\Lambda}}_m\}_{m=1}^M$ in the approximate posterior.

We shall briefly describe ways to apply the proposed model to other alignment methods. For applying the unified model to existing alignment models that use intra-recording duration model, it suffices to modify the state transition dynamics from Eq. 1 to Eq. 6. Applying the switching-state unified model may require derivation of a structured mean-field inference [26] that decouples μ and u.

5. EVALUATION

Here, we compare the absolute error percentile of alignment obtained from different duration models, to assess the effectiveness of the switching-state intra-recording duration model and the unified duration model.

5.1. Experimental condition

We evaluated our duration model by evaluating the alignment error percentile when aligning Chopin's *Mazurka*. We chose pieces for which reliable ground truth "reverse conducting" data were available [12], which yielded in nine pieces, each with 2 to 5 recordings. For each song, we compared the estimated alignment with a human-annotated data. We computed the chroma and the delta-chroma feature using a sampling frequency of 44.1kHz at 25 frames per second. For the unified duration model, m was set to 1 and α was set to 0.1. These parameters moderately encourages the relative tempo to be near 1. The prior precision of a_n, ι, was set to 0.01, allowing for standard deviation of about three frames. The prior precision of l_{nd}, λ_0, was set to 30. The hyperparameters to Λ_0 was set to $n_0 = I$ and $\boldsymbol{W}_0 = 100\boldsymbol{I}_d$. For the switching-state intra/unified duration model, we set $\xi_0 = 0.1$, which encourages a sparse transition structure. The number of covariance matrices to use, M, was set to 5. For parameters related to alignment but not relevant for the duration model, we used the parameters listed in [19]. We compared our method against six different methods:

DTW Path optimization approach based on minimizing the total distance (cosine distance) using path constraint in [17].

Baseline The method introduced in Section 3.3, using a geometric duration pdf, *i.e.*, treating z in Section 3.3 like a HMM.

Inter Using Eq. 4 as the duration model, using a single spherical covariance matrix for the innovation $\epsilon_n^{(\mu)}$.

Intra Using Eq. 1 as the dynamics model, using a single spherical covariance matrix for the innovation $\epsilon_n^{(\mu)}$. It is thus similar in spirit to existing intra-recording duration models [6, 20].

Switching-state (Intra) Same as method **Intra**, except we used a switching-state dynamics by setting $M = 5$.

Switching-state (Unified) The proposed method.

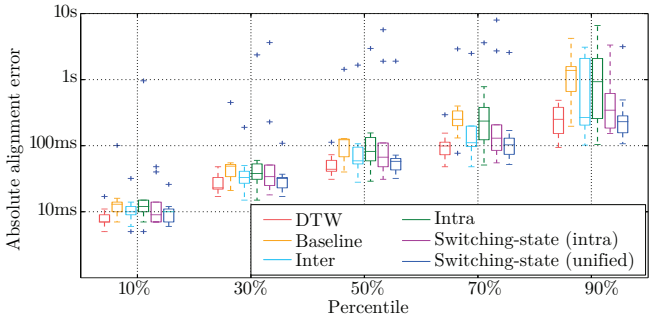

Figure 4: Absolute alignment error percentile.

5.2. Result and discussion

The result is presented in Fig. 4. We can confirm that, by comparing the baseline and method **inter** or **intra**, both inter- and intra-recording duration models are useful in probabilistic audio alignment. It seems that in the given dataset, inter-recording duration model is better than the intra-recording model. This is perhaps because the baseline method sometimes segments the recordings in musically irrelevant ways. For example, the method may segment a single note into the attack, the decay and the sustain. Intra-recording models are meaningless in such a segmentation, making inter-recording model more relevant.

Furthermore, by comparing **intra** and **switching-state (intra)**, we see that switching-state model is more effective than the non-switching counterpart. This suggests that there are indeed portions in which the tempo innovation changes drastically.

Furthermore, we note that the unified duration model provides additional improvement, by comparing **inter**, **switching-state (intra)** and **switching-state (unified)**. This suggests that the unified model is capable of utilizing the strengths of both methods. The decrease in the variance of the estimates (range of the whiskers) in the switching-state unified model is also indicative of the applicability of the proposed model to a wider variety of songs.

Finally, switching-state unified model performs comparably to DTW, though DTW has fewer outliers that consistently fail to align. The outliers occur in a few specific pieces regardless of the duration model; this suggests that the problem is rooted in the baseline alignment method. Specifically, the baseline method seems to fail when the balance of note dynamics within a chord vary significantly, suggesting the need for a better spectral model. Furthermore, segmenting the input audio segments into N segments, the fundamental modeling idea behind the baseline method, is a difficult problem, more so since the nth segment of different signal should describe a same position in the underlying piece of music.

6. CONCLUSION

This paper presented a duration model for probabilistic multiple audio alignment. The proposed method integrates inter- and intra-recording duration models, and allows the tempo trajectory to vary with sporadically changing covariance matrices. Evaluation demonstrated that the unification of inter- and intra-recording models is effective and that allowing the tempo curve to vary sporadically is effective. Future work includes application of the proposed method to wider problem domains, and the improvement of the baseline alignment method.

7. REFERENCES

[1] S. Ewert S. Wang and S. Dixon, "Robust joint alignment of multiple versions of a piece of music," in *ISMIR*, 2014, pp. 83–88.

[2] S. Dixon and G. Widmer, "MATCH: A music alignment tool chest," in *ISMIR*, 2005.

[3] M. Müller and S. Ewert, "Towards timbre-invariant audio features for harmony-based music," *TASLP*, vol. 18, no. 3, pp. 649–662, Mar. 2010.

[4] M. Grachten, M. Gasser, A. Arzt, and G. Widmer, "Automatic alignment of music performances with structural differences," in *ISMIR*, November 2013.

[5] B. Niedermayer and G. Widmer, "A multi-pass algorithm for accurate audio-to-score alignment," in *ISMIR*, 2010, pp. 417–422.

[6] C. Raphael, "A hybrid graphical model for aligning polyphonic audio with musical scores," in *ISMIR*, 2004, pp. 387–394.

[7] A. Cont, "A coupled duration-focused architecture for real-time music-to-score alignment," *PAMI*, vol. 32, no. 6, pp. 974–987, 2010.

[8] Z. Duan and Bryan P., "A state space model for online polyphonic audio-score alignment," in *ICASSP*, 2011, pp. 197–200.

[9] T. Otsuka, K. Nakadai, T. Ogata, and H. G. Okuno, "Incremental Bayesian audio-to-score alignment with flexible harmonic structure models," in *ISMIR*, 2011, pp. 525–530.

[10] S. Sako, R. Yamamoto, and T. Kitamura, "Ryry: A real-time score-following automatic accompaniment playback system capable of real performances with errors, repeats and jumps," in *AMT*, 2014, pp. 134–145.

[11] C. Joder, S. Essid, and G. Richard, "A conditional random field framework for robust and scalable audio-to-score matching," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 8, pp. 2385–2397, Nov 2011.

[12] C. S. Sapp, "Comparative analysis of multiple musical performances," in *ISMIR*, 2007, pp. 2–5.

[13] D. Stowell and E. Chew, "Maximum a posteriori estimation of piecewise arcs in tempo time-series," in *From Sounds to Music and Emotions*, LNCS(7900), pp. 387–399. Springer, 2013.

[14] V. Konz, *Automated methods for audio-based music analysis with applications to musicology*, Ph.D. thesis, Saarland University, 2012.

[15] R. Hennequin, J. J. Burred, S. Maller, and P. Leveau, "Speech-guided source separation using a pitch-adaptive guide signal model," in *ICASSP*, 2014, pp. 6672–6676.

[16] A. Maezawa and H. G. Okuno, "Audio part mixture alignment based on hierarchical nonparametric Bayesian model of musical audio sequence collection," in *ICASSP*, 2014, pp. 5212–5216.

[17] R. B. Dannenberg and N. Hu, "Polyphonic audio matching for score following and intelligent audio editors," in *ICMC*, Sept. 2003.

[18] N. Montecchio and A. Cont, "A unified approach to real time audio-to-score and audio-to-audio alignment using sequential Montecarlo inference techniques," in *ICASSP*, 2011, pp. 193–196.

[19] A. Maezawa, K. Itoyama, K. Yoshii, and H.G. Okuno, "Bayesian audio alignment based on a unified generative model of music composition and performance," in *ISMIR*, 2014.

[20] N. Montecchio and A. Cont, "A unified approach to real time audio-to-score and audio-to-audio alignment using sequential Montecarlo inference techniques," in *ICASSP*, 2011, pp. 193–196.

[21] P. Desain and H. Honing, "Does expressive timing in music performance scale proportionally with tempo?," *Psychological Research*, vol. 56, no. 4, pp. 285–292, 1994.

[22] Y. Gu and C. Raphael, "Modeling piano interpretation using switching Kalman filter," in *ISMIR*, October 8-12 2012.

[23] A.T. Cemgil, B. Kappen, P. Desain, and H. Honing, "On tempo tracking: Tempogram representation and Kalman filtering," *J. New Music Research*, vol. 29, no. 4, pp. 259–273, 2000.

[24] T. Otsuka, T. Takahashi, H. G. Okuno, K. Komatani, T. Ogata, K. Murata, and K. Nakadai, "Incremental polyphonic audio to score alignment using beat tracking for singer robots," in *IROS*, 2009, pp. 2289–2296.

[25] M. J. Beal, *Variational algorithms for approximate Bayesian inference*, Ph.D. thesis, University College London, 2003.

[26] Z. Ghahramani and G. E. Hinton, "Variational learning for switching state-space models," *Neural computation*, vol. 12, no. 4, pp. 831–864, 2000.