POLYPHONIC AUDIO-TO-SCORE ALIGNMENT BASED ON BAYESIAN LATENT HARMONIC ALLOCATION HIDDEN MARKOV MODEL

Akira Maezawa,[†] *Hiroshi G. Okuno,*[†] *Tetsuya Ogata,*[†] *Masataka Goto*[‡]

[†]Dept. of Intelligence Science and Technology
Graduate School of Informatics, Kyoto University
Sakyo, Kyoto 606-8501 Japan

[‡]National Institute of Advanced Industrial
Science and Technology (AIST)
Tsukuba, Ibaraki 305-8568 Japan

ABSTRACT

This paper presents a Bayesian method for temporally aligning a music score and an audio rendition. A critical problem in audio-to-score alignment is in dealing with the wide variety of timbre and volume of the audio rendition. In contrast with existing works that achieve this through ad-hoc feature design or careful training of tone models, we propose a Bayesian audio-to-score alignment method by modeling music performance as a Bayesian Hidden Markov Model, each state of which emits a Bayesian signal model based on Latent Harmonic Allocation. After attenuating reverberation, variational Bayes method is used to iteratively adapt the alignment, instrument tone model and the volume balance at each position of the score. The method is evaluated using sixty works of classical music of a variety of instrumentation ranging from solo piano to full orchestra. We verify that our method improves the alignment accuracy compared to dynamic time warping based on chroma vector for orchestral music, or our method employed in a maximum likelihood setting.

Index Terms— Audio-to-score alignment, Variational Bayes inference

1. INTRODUCTION

Audio-to-score alignment, the task of finding a temporal mapping between a music score and audio signal, is a critical task in music information retrieval. It is required whenever a system needs to coordinate a music score (e.g. a standard MIDI file) and an audio signal, such as score-informed source separation [1, 2], or score-informed analysis of music performance [3, 4, 5, 6, 7].

Audio-to-score alignment is difficult because the music score conveys only the sequence of instrument and pitch in which a piece should be played – very little on the fluctuation of tempo, timbre and balance of dynamics are conveyed. For example, an orchestra might emphasize a particular instrument part more than the other compared to another orchestra. One orchestra might have a brilliantly-sounding brass section compared to another. One orchestra might slow down a particular section of music, while another orchestra might speed it up a bit. Audio-to-score alignment should absorb such discrepancies in dynamics timbre and tempo.

Most approaches are based on designing timbre-robust feature for alignment, using ad-hoc distance measure. We found that, while there are basic tenets of designing features and distance measures, a big portion of it is still an art that requires careful adjustments. A critical problem is that of removing timbral dependence; aligning a piece of music that is rich in timbre, such as a symphony, is difficult.

We believe that the problem of spectral mismatch could be alleviated only to an extent through design of timbre-robust features. To cope with timbre, we should instead consider timbre as an unknown quantity *a priori*, that we infer from the observed musical audio. Such philosophy calls for a Bayesian treatment of timbre in audio-to-score alignment.

In this paper, we propose an audio-to-score alignment method by fitting a Bayesian Hidden Markov Model (HMM) based on Latent Harmonic Allocation (LHA) signal model [8]. LHA is a Bayesian treatment of musical signal based on fitting a superposition of histograms onto the observed power spectrum. Our method is based on three assumptions. First, we assume that the *volume* of each note is different, but remains more-or-less consistent *within* a note (intra-note volume consistency). Second, the *timbre* of each pitch / instrument pair is unique, but remains consistent throughout a piece (intra-music timbre consistency). Third, each position of the music score emits a combination of pitch / instrument pairs and a stationary noise. We evaluate the effectiveness of treating timbre and volume as random quantities using sixty pieces of classical music.

2. EXISTING WORKS

Audio-to-score of a piece with a wide range of timbre has been attacked mainly using two approaches to feature selection. First approach uses features that are robust to timbral differences. For example, many studies use chroma vector, a timbre-robust feature that reflects the power in a particular pitch-class [9, 10, 11, 12, 13]. Other studies use features that are reminiscent of the chroma vector, but with improved robustness against timbral variety [14, 15]. Second approach is based on a generative model of spectrograms [16, 17]. For temporal matching, Dynamic Time Warping (DTW), HMM, or Dynamic Bayesian Model (DBM) have been employed, though DTW is used, by far, most frequently.

3. BAYESIAN AUDIO-TO-SCORE ALIGNMENT BASED ON LHA-HMM

Our method estimates audio-to-score alignment by modeling the musical audio signal as a Bayesian HMM, each state of which emits a signal based on LHA model.

3.1. Preprocessing

Given an input signal $o(t)$, its short-time Fourier Transform (STFT), $O(f,t)$, of size F-by-T is evaluated. Then, reverberation is attenuated by modeling $O(f,t)$ as an auto-regressive process with $X(f,t)$ as the source signal. This model was inspired from that used in speech dereverberation [18]. Each frequency bin contains linear predictive coefficients $G(f,i)$ of order P. In other words, we model the dereverberated STFT $X(f,t)$ as follows:

$$O(f,t) = X(f,t) + \sum_{i}^{P} O(f,t-i)G(f,i) \qquad (1)$$

$G(f,i)$ is estimated using linear prediction.

3.2. Latent Harmonic Allocation

LHA is a generative signal model proposed by Yoshii, which interprets the spectrogram as a histogram, each bin of which is considered

to be multiple draws from various musical instrument [8]. The probability of observing the energy from a particular musical instrument follows a Normal distribution centered around integer multiple of its fundamental frequency. The probability of drawing a particular overtone of a particular instrument, both of which follow multinomial distribution, have a conjugate prior. Variational expectation-maximization is used for inference.

3.3. Joint Estimation of Alignment, Timbre, Pitch and Volume

We assume that power contained in each time-frequency bin of $X(f,t)$ originated from some combination of power originating from different musical instruments playing at different pitch. Hence, the source signal $X(f,t)$ is modeled as a sum of music instrument histograms. Assuming that a particular music contains I unique instrument/pitch pairs, each with the fundamental frequency distributed normally with mean μ_i and precision λ_i, the model fits a sequence of a combination of these pairs onto $X(f,t)$.

To attain intra-music timbre consistency for each of I instrument/pitch pairs, each instrument/pitch pair has a unique *tone-model*, consisting of the fundamental frequency and the relative strengths of the overtones. This tone-model is shared for all occurrences of the given instrument/pitch pair in a piece. Our model models intra-note volume consistency by incorporating a I-dimensional *note emission* vector, ith element of which indicates the likelihood that ith instrument/pitch pair contributes to the observation.

The model joins a sequence of D note-emission vectors by a left-to-right HMM, where D is the number of states of the music score. The emission vector associated with the dth state is set such that it is highly likely to emit pitch/instrument pairs that are notated at that point in the music score, and highly unlike to emit others. $s_d(t)$ is a binary vector, which is 1 if state at time t is d, and 0 otherwise.

To infer the alignment, we introduce various latent variables. $Y = \{Y(f,i)\}$ is a random variable of size $I \times F$, each of which is a J-dimensional multinomial variable jth element of which governs the ratio of power the jth overtone contributes in frequency f for ith instrument/pitch pair. Furthermore, $Z = \{Z(f,d)\}$ is a random variable of size $F \times D$, each of which is a I-dimensional multinomial variable, ith element of which governs the ratio of power the the ith instrument/pitch pair contributes in frequency f.

The complete data likelihood is given as follows:

$$p(X,Z,Y,S,E,A,\mu,\lambda) = p(X|S,Y,Z,\mu,\lambda)p(Z|E,S) \times p(Y|Z,S,A)p(S|\tau)p(E)p(A)p(\tau)p(\mu,\lambda) \quad (2)$$

where

$$p(X|Y,Z,\mu,\lambda,S) = \prod_{t,i,j,f,d} \mathcal{N}\left(f|j\mu_i, j^2\lambda_i^{-1}\right)^{X(f,t)Z_i(f,d)Y_j(f,i)S_d(t)}$$

$$p(Z|E,S) = \prod_{t,i,f,d} e_i(d)^{X(f,t)Z_i(f,d)S_d(t)}$$

$$p(Y|A,Z,S) = \prod_{t,i,j,f,d} a_j(i)^{X(f,t)Z_i(f,d)Y_j(f,i)S_d(t)}$$

$$sp(\mu,\lambda;\nu,b,m) = \prod_k \mathcal{NG}(\mu_k,\lambda_k;m_k,b_k,l_k,\nu_k)$$

$$p(\tau) = \prod_j \mathrm{Dir}(\tau(j);\tau_0(j)) \qquad p(S|\tau) = \prod_t \prod_{j,k} \tau_j(k)^{s_j(t)s_k(t-1)}$$

$$p(E;\epsilon) = \prod_d^D \mathrm{Dir}(e(d);\epsilon_d) \qquad p(A;\alpha) = \prod_k \mathrm{Dir}(a(k);\alpha(k))$$

$\mathcal{N}(f|\mu,\lambda^{-1})$ is a discrete distribution over $f \in Z$, in which:

$$\mathcal{N}(f|\mu,\lambda^{-1}) = \int_{f-1/2}^{f+1/2} \mathcal{N}_c(\tilde{f}|\mu,\lambda^{-1})d\tilde{f}$$

where $\mathcal{N}_c$ is a Normal distribution. We approximate this $\mathcal{N}$ by

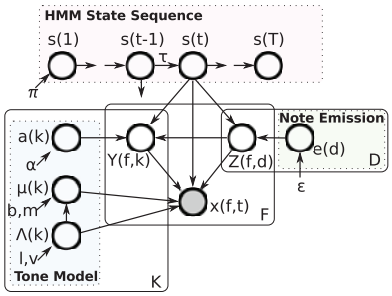

Fig. 1. Graphical model of our method.

$\mathcal{N}(f) \approx \int_{f-1/2}^{f+1/2} \mathcal{N}_c(f)d\tilde{f} = \mathcal{N}_c(f)$. $\mathcal{NG}$ is a Normal-Gamma distribution, whose probability density is given as follows:

$$\mathcal{NG}(\mu,\lambda|m,b,l,\nu) = \frac{\nu^l \lambda_k^{l-1/2}\sqrt{b}}{\Gamma(l)\sqrt{2\pi}} \exp\left(-\frac{1}{2}b\lambda(\mu-m)^2 - \lambda\nu\right) \quad (3)$$

where $\Gamma(i)$ is the Gamma function. $\mathrm{Dir}(a)$ is the dirichlet distribution. The complete graphical model is shown in Fig. 1.

We seek to maximize the posterior $p(S,Z,Y,E,A,\mu,\lambda,\tau|X)$, but this is analytically intractable. Therefore, we instead maximize an approximation of the posterior, $q(S,Z,Y,E,A,\mu,\lambda,\tau)$. We assume the approximation is factorized into the following form:

$$q(S,Z,Y,E,A,\mu,\lambda,\tau) = q_S(S)q_Z(Z)q_Y(Y)q_E(E)q_A(A)q_{\mu,\lambda}(\mu,\lambda)q_\tau(\tau) \quad (4)$$

We optimize the approximate posterior by minimizing the Kullback-Leibler divergence between it and the posterior. Such variational minimization yields in the following update [19]:

$$\log q_{Z_i}(Z_i) := c + \langle \log p(X,S,Z,Y\cdots\tau)\rangle_{\neg Z_i} \quad (5)$$

$\neg Z_i$ means all factors of $q(S,Z,Y,\cdots,\tau)$ excluding Z_i, and $\langle f(x,y)\rangle_y$ means to take the expectation of $f(x,y)$ under $q(y)$. The iteration terminates when the following quantity has converged:

$$\mathcal{F}(q_x(x),\theta) = \int q_Z(Z)\log\frac{p(X,Z)}{q_Z(Z)}dZ \quad (6)$$

For convenience, we define the following quantities:

$$\xi_j(f,i) = \langle Y_j(f,i)\rangle_Y \quad (7) \qquad \eta_d(i) = \langle s_d(t)\rangle_s \quad (10)$$

$$\gamma_i(f,d) = \langle Z_i(f,d)\rangle_Z \quad (8) \qquad lA_j(i) = \langle \log a_j(i)\rangle_E \quad (11)$$

$$l\tau_{d'}(d) = \langle \log \tau_{d'}(d)\rangle_\tau \quad (9) \qquad lE_i(d) = \langle \log e_i(d)\rangle_E \quad (12)$$

$$l\mathcal{N}_f(i,j) = \left\langle \log \mathcal{N}\left(f|j\mu_i, j^2\lambda_i^{-1}\right)\right\rangle_{\mu,\lambda} \quad (13)$$

$$\zeta_i(t,j) = \langle s_i(t)s_j(t-1)\rangle_s \quad (14)$$

where

$$lA_j(i) = \psi(\alpha_j(i)) - \psi\left(\textstyle\sum_{l=1}^M \alpha_l(i)\right) \quad (15)$$

$$lE_i(d) = \psi(\epsilon_i(d)) - \psi\left(\textstyle\sum_{l=1}^K \epsilon_l(d)\right) \quad (16)$$

$$l\mathcal{N}_f(i,j) = -\frac{1}{2}\left(\frac{l_i}{\nu_i}\left(\frac{f}{j} - m_i\right)^2 + \frac{1}{b_i}\right) - \log 2\pi\nu_i + \psi(l_i) \quad (17)$$

Here, $\psi(x)$ is the digamma function.

3.4. Updating Note Emission

First we update the responsibility ith pitch/instrument pair has at a given frequency f, for each state of the HMM d to the following:

$$q_Z(Z) = \prod_{i,f,d} \gamma_i(f,d)^{Z_i(f,d)} \text{ where } \gamma_i(f,d) = \frac{\rho_i(f,d)}{\sum_i \rho_i(f,d)} \quad (18)$$

where ρ is the following:

$$\log \rho_i(f,d) = \Big(\sum_t X(f,t)\eta_d(t) \Big) \times \Big[lE_i(d) + \sum_j \xi_j(f,i)\big(l\mathcal{N}_f(i,j) + lA_j(i)\big) \Big] \quad (19)$$

Furthermore, the note-emission vector is updated as follows:

$$q_E(E) = \prod_i \mathrm{Dir}\Big(e_d; \epsilon_i(:) + \sum_{t,f,j} \eta_{:}(t)X(f,t)\gamma_{i,j}(f,:)\Big) \quad (20)$$

3.5. Updating Tone-model

First, we update the responsibility jth overtone has at a given frequency f, for each note / instrument pair i:

$$q_Y(Y) = \prod_{j,i,f} \xi_j(i,f)^{Y_j(f,i)} \text{ where } \xi_j(i,j) = \frac{\phi_j(f,i)}{\sum_k \phi_k(f,i)} \quad (21)$$

where ϕ the following:

$$\log \phi_j(f,i) = \Big(l\mathcal{N}_f(i,j) + \log a_j(i) \Big) \times \sum_{t,d} X(f,t)\eta_d(t)\gamma_i(f,d) \quad (22)$$

The fundamental frequency and its variance is set to the following:

$$q_{\mu,\lambda}(\mu,\lambda) = \prod_k \mathcal{NG}(\mu_k, \lambda_k; m_k, b_k, l_k, \nu_k) \quad (23)$$

with

$$m_k := \frac{m_k b_k + N_{\gamma,k} \langle x/j \rangle_{\psi_k}}{b_k + N_{\gamma,k}} \quad (24)$$

$$b_k := b_k + N_{\gamma,k} \quad (25)$$

$$l_k := l_k + N_{\gamma,k} \quad (26)$$

$$\nu_k := \nu_k + \frac{1}{2}\frac{b_k N_{\gamma,k}}{b_k + N_{\gamma,k}} \Bigg(\Big(\langle x/j \rangle_{\psi(k)} - m_k \Big)^2 + N_{\gamma,k}^2 \Big\langle x/j - \langle x/j \rangle_{\psi(k)} \Big\rangle^2_{\psi(k)} \Bigg) \quad (27)$$

where $\psi_k(f,j)$ is the following multinomial distribution:

$$\psi_{f,j}(k) = \sum_{d,t} \gamma_k(f,d)\xi_j(f,k)\eta_{(}d,t)X(f,t)/N_{\gamma,k} \quad (28)$$

and

$$N_{\gamma,k} = \sum_{f,j}\sum_{d,t} \gamma_k(f,d)\xi_j(f,k)\eta_{(}d,t)X(f,t) \quad (29)$$

The relative strength of each overtone is updated to the following:

$$q_A(A) = \prod_i \mathrm{Dir}\left(a(i); \alpha(i) + \sum_{t,f,d} X(f,t)\gamma_i(f,d)\eta_d(t)\xi(f,i) \right) \quad (30)$$

3.6. Updating HMM

The state transition probability τ is updated as follows:

$$q_\tau(\tau) = \prod_k \mathrm{Dir}\left(\tau(k); \sum_t \zeta(t,k) + \tau_0(k) \right) \quad (31)$$

The probability of state sequence is given as follows:

$$q_S(S) = \prod_{t=1}^{T} \prod_{d,d'} \big(\langle \log \tau_j(i) \rangle_\tau^{S_{d'}(t-1)} \prod_f \kappa_d(f)^{X(f,t)} \big)^{S_d(t)} \quad (32)$$

where

$$\log \kappa_d(f) = \sum_i \gamma_i(f,d)\Big(lE_i(d) + \sum_j \xi_j(f,i)\big(l\mathcal{N}_f(i,j) + lA_j(i)\big) \Big) \quad (33)$$

Note that this has the same functional form as a HMM, with transition probability replaced by exponent of log-expectation, and the emission probability replaced $\kappa_d(f)$. Hence, $\kappa_d(f)$ can be considered as a subnormalized histogram that represents the spectrum the dth state is likely to emit. To compute the expectation of the sufficient statistics η and ξ, forward-backward algorithm is used. Forward probability α and backward probability β are given as the following recursions:

$$\alpha(t,d) = p(s(t)|X(1)\cdots X(t)) = \frac{1}{Z}\sum_{d'} \Big(\alpha(t-1,d')e^{l\tau_{d'}(d)} \Big) \prod_f \kappa_d(f)^{X(f,t)} \quad (34)$$

$$\beta(t,d) = p(X_{t+1}(f)\cdots X_T(f)|s_d(t)=1) = \sum_{d'} \beta(t+1,d') \prod_f \kappa_{d'}(f)^{X(f,t+1)} e^{l\tau_d(d')} \quad (35)$$

These are used to arrive at the sufficient statistics:

$$\langle s_j(t) \rangle_s = \frac{1}{Z}\alpha(t,j)\beta(t,j) \quad (36)$$

$$\langle s_j(t-1)s_k(t) \rangle_s = \frac{\alpha(t-1,j)\beta(t,k)}{Z} \prod_f \kappa_k(f)^{X(f,t)} e^{l\tau_j(k)} \quad (37)$$

4. EXPERIMENT

We evaluate the performance of our system against a widely-used alignment method. Moreover, we evaluate the significance of performing audio-to-score alignment in a Bayesian setting as opposed to maximum likelihood estimation, and the effectiveness of adapting the volume and timbre information for aligning music.

Our method is tested against three conditions. First condition is alignment based on Dynamic Time-warping (DTW) of chroma-vector, perhaps the most popular method for audio-to-score alignment. Second, alignment based on non-adaptive, maximum-likelihood LHA-HMM is used as a baseline for non-Bayesian framework of our method. Essentially, Viterbi decoding is done on a HMM which has $\langle \tau \rangle_{q_\tau(\tau)}$, instead of $\exp \langle \log \tau \rangle_{q_\tau(\tau)}$ for state transition probability. Furthermore, κ is replaced with a fixed and normalized spectral histogram $\kappa_d(f) = \sum_{i,j} \langle \epsilon_i(d) \rangle \langle \alpha_j(i) \rangle \mathcal{N}(f|j\mu_i, l_k/(b_k v_k))$; this signal model is similar to one used in [16]. Finally, LHA-HMM without volume/timbre adaptation is used as a baseline for Bayesian LHA-HMM without timbre update.

For evaluation, we align sixty pieces from RWC Classical Music Database [20], and compare them against the database's beat data [21]. To evaluate our method for classical music with various instrumentations, we divide the database into five categories: orchestral (C001 to C010), non-orchestral ensemble (C012 to C021), solo piano (C022 to C035), solo instrumental/duo (C036-C011, C011), and vocal music (C045 to C050). C005 is omitted because it includes a cadenza, a freely-played section that is not notated on the score.

All signals are downsampled to 8kHz and analyzed using Hanning window of length 2048 with overlap of 400 samples. One persistent tone with fundamental of 500Hz and variance of each Gaussian set to 500Hz was used as the noise model. The prior of note emission are set such that, all unnotated notes have hyperparameter set to 10^{-6}, and all notated notes and noise set to 1 (non-informative prior). For each tone model, m_k was set to the notated fundamental frequency, and the variance to 50Hz ($l_k = 10^6$, $\nu_k = 50 \times 10^6$, $b_k = 50 \times 10^6$). Moreover, we consider up to the tenth overtone ($J = 10$) for each tone-model, and the J-dimensional timbre model $\alpha_j(k)$ set to non-informative.

To optimize the posterior, we iteratively update γ, ξ, τ and η until convergence, after which A is updated. This is repeated until convergence. After convergence E was updated. μ and λ were not updated, as we assumed the shape of harmonics do not change significantly within a piece of music. The result is shown in Table 1.

Table 1. Percentage of estimated alignment whose error ϵ is within a given margin. "Type" indicates the category of pieces, where OR is orchestral, EN is non-orchestral ensemble, PN is piano-solo, IN is non-piano solo/duo, and VO is vocal music. Four conditions, Chroma-based DTW, Maximum Likelihood of our method, our method without timbre and volume update, and proposed method, are labeled "C," "ML," "N," and "P," respectively.

Type		ϵ <50ms	ϵ <.1s	ϵ <.5s	ϵ < 1s
OR	C	8.0%	16.3%	57.7%	70.3%
OR	ML	1.2%	1.8%	4.1%	6.5%
OR	N	24.4%	41.4%	65.1%	72.4%
OR	P	**28.2%**	**46.8%**	**70.8%**	**75.0%**
EN	C	14.7%	29.1%	69.5%	**83.4%**
EN	ML	0.1%	0.1%	0.5%	1.1%
EN	N	**16.8%**	**36.0%**	71.7%	80.7%
EN	P	16.2%	35.5%	**72.0%**	80.9%
PN	C	**12.3%**	**26.4%**	56.3%	**65.4%**
PN	ML	0.1%	0.1%	0.5%	1.0%
PN	N	6.7%	18.8%	56.3%	63.3%
PN	P	6.6%	18.8%	**56.9%**	63.7%
IN	C	11.8%	25.1%	**71.1%**	**81.8%**
IN	ML	1.7%	2.6%	3.4%	4.1%
IN	N	24.3%	40.4%	67.2%	73.2%
IN	P	**24.8%**	**40.7%**	67.5%	73.4%
VO	C	6.2%	11.9%	39.1%	**54.3%**
VO	ML	0.1%	0.1%	0.2%	0.2%
VO	N	**8.3%**	**14.4%**	**40.2%**	49.5%
VO	P	7.9%	13.9%	39.6%	49.2%

5. DISCUSSION

Our method is capable of generating alignment whose quality is comparable to that of well-established chroma-based DTW alignment. This is significant because DTW based on timbre-robust features is still an art that requires finesse in design of features and dissimilarity measure. On the other hand, the only ad-hoccery employed in our method was the choice of the magnitude of note-emission vector for un-notated notes, and the initial variance of each Normal-Gamma distribution. Otherwise, our method was initialized with non-informative priors. Treating timbre and volume as fixed quantity ("ML") significantly degrades the alignment, suggesting that it is in the Bayesian treatment of timbre and volume that achieves the performance of our method.

The performance of our model tends to degrade slightly when we update the timbre and volume, as seen in the results of "N" and "P." This is because posterior of a Dirichlet (i.e. timbre and note emission) has smaller variance. Hence, updating timbre and volume model makes our model more selective to musical signals that satisfy intra-note volume consistency and intra-music timbre consistency. In reality, these assumptions do not strictly hold, meaning that it is better to deal with timbre and note emission as Dirichlet with a large variance instead of coercing it into a particular multinomial. It is important to note that both "N" and "P" are guaranteed to converge; they are two different ways to align using LHA-HMM.

Our method performs poorly in piano solo and vocal music, as seen from the low percentage of errors that lie within 50ms. This is because they violate intra-note volume consistency and intra-music timbral consistency, respectively. Intra-note volume consistency is violated in piano music because the volume of piano decays within a note. Intra-music timbre consistency is violated in vocal music because a singer can produce a variety of phonemes.

On the other hand, the method is effective for orchestral pieces. This is because orchestral music is rich in timbral content and dynamic range; achieving timbre robustness is hard in this situation, but our model absorbs differences in timbre and dynamics.

6. CONCLUSION

This paper presented a Bayesian approach to audio-to-score alignment using LHA model and HMM. Our method, unlike most existing methods using DTW, does not require careful adjustments of model parameters and skillful design of timbre-robust features and dissimilarity measures. Yet, our method performed similarly to that of DTW-based method using ad-hoc timbre-robust features.

In the future, we plan to incorporate more sophisticated model for note duration. Moreover, we plan to explore other models for timbre and dynamics, to better deal with piano and vocal music.

Acknowledgment The authors thank Kazuyoshi Yoshii for helpful discussions on Latent Harmonic Allocation inference. This work was funded by Kakenhi(S) and GCOE.

7. REFERENCES

[1] K. Itoyama et al., "Integration and Adaptation of Harmonic and Inharmonic Models for Separating Polyphonic Musical Signals," in *ICASSP*, Apr. 2007, pp. I–57–I–60.

[2] Y. Han and C. Raphael, "Desoloing Monaural Audio Using Mixture Models," in *ISMIR*, 2007, pp. 145–148.

[3] A. Maezawa et al., "Query-By-Conducting: An Interface to retrieve classical-music interpretations by real-time tempo input," in *ISMIR*, 2010, pp. 477–482.

[4] A. Maezawa et al., "Bowed String Sequence Estimation of a Violin Based on Adaptive Audio Signal Classification and Context-Dependent Error Correction," in *ISM*, 2009, pp. 9–16.

[5] A. Maezawa et al., "Violin Fingering Estimation Based on Violin Pedagogical Fingering Model Constrained by Bowed Sequence Estimation from Audio," in *IEA/AIE*, 2010.

[6] M. Molina-solana et al., "Evidence for pianist-specific rubato style in chopin nocturnes," in *ISMIR*, 2010, pp. 225–230.

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

[8] K. Yoshii and M. Goto, "Infinite Latent Harmonic Allocation: A nonparametric Bayesian approach to multipitch analysis," in *ISMIR*, 2010, pp. 309–314.

[9] C. Joder et al., "An improved hierarchical approach for music-to-symbolic score alignment," in *ISMIR*, 2010, pp. 39–44.

[10] R. Macrae and S. Dixon, "Accurate real-time windowed time warping," in *ISMIR*, 2010, pp. 423–428.

[11] B. Niedermayer et al., "A multi-pass algorithm for accurate audio-to-score alignment," in *ISMIR*, 2010, pp. 417–422.

[12] N. Orio et al., "Score Following : State of the Art and New Developments," in *NIME*, 2003, pp. 36–41.

[13] N. Hu et al., "Polyphonic audio matching and alignment for music retrieval," in *WASPAA*, 2003, pp. 185–188.

[14] M. Muller and S. Ewert, "Towards Timbre-Invariant Audio Features for Harmony-Based Music," *IEEE TASLP*, vol. 18, no. 3, pp. 649–662, Mar. 2010.

[15] M. Muller and F. Kurth, "Enhancing similarity matrices for music audio analysis," in *ICASSP*, 2006, pp. 9–12.

[16] C. Raphael, "A Hybrid Graphical Model for Aligning Polyphonic Audio with Musical Scores," in *ISMIR*, 2004, pp. 387–394.

[17] A.T. Peeling, P. Cemgil and S. Godsill, "A Probabilistic Framework for Matching Music Representations," in *ISMIR*, 2007, pp. 267–272.

[18] T. Yoshioka et al., "Dereverberation by using time-variant nature of speech production system," *EURASIP J. Adv. Signal Process*, vol. 2007, no. 2, pp. 6–6, 2007.

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

[20] M. Goto, "Development of the RWC Music Database," in *Int'l Congress on Acoustics*, 2004, vol. I, pp. 553–556.

[21] M. Goto, "AIST Annotation for the RWC Music Database," in *ISMIR*, 2006, pp. 359–360.