INCREMENTAL BAYESIAN AUDIO-TO-SCORE ALIGNMENT WITH FLEXIBLE HARMONIC STRUCTURE MODELS

Takuma Otsuka[†], Kazuhiro Nakadai[‡], Tetsuya Ogata[†], and Hiroshi G. Okuno[†]

[†] Graduate School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501 Japan
{ohtsuka, ogata, okuno }@kuis.kyoto-u.ac.jp

[‡] Honda Research Institute Japan, Co., Ltd.
Wako, Saitama 351-0114, Japan
nakadai@jp.honda-ri.com

ABSTRACT

Music information retrieval, especially the audio-to-score alignment problem, often involves a matching problem between the audio and symbolic representations. We must cope with uncertainty in the audio signal generated from the score in a symbolic representation such as the variation in the timbre or temporal fluctuations. Existing audio-to-score alignment methods are sometimes vulnerable to the uncertainty in which multiple notes are simultaneously played with a variety of timbres because these methods rely on static observation models. For example, a chroma vector or a fixed harmonic structure template is used under the assumption that musical notes in a chord are all in the same volume and timbre. This paper presents a particle filter-based audio-to-score alignment method with a flexible observation model based on latent harmonic allocation. Our method adapts to the harmonic structure for the audio-to-score matching based on the observation of the audio signal through Bayesian inference. Experimental results with 20 polyphonic songs reveal that our method is effective when more number of instruments are involved in the ensemble.

1. INTRODUCTION

Music information retrieval tasks require a robust inference under the uncertainty in musical audio signals. For example, a polyphonic or multi-instrument aspect encumbers the fundamental frequency estimation [10, 15] or instrument identification [9]. Overcoming the uncertainty in musical audio signals is a key factor in the machine comprehension of musical information. The audio-to-score alignment technology shares this uncertainty problem in that an audio signal performed by human musicians has a wide range of varieties given a symbolic score due to the musicians' expressiveness. For example, the type of instruments and the temporal or pitch fluctuations affect the resulting audio signals.

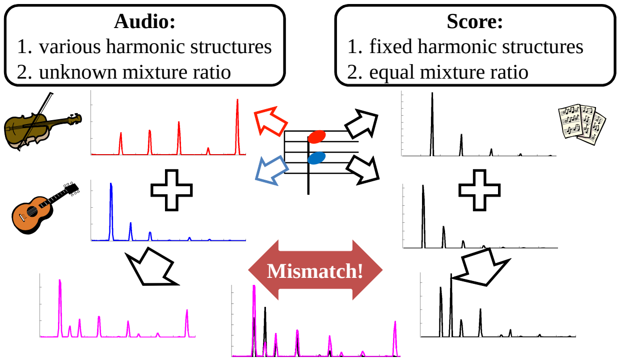

Figure 1. The issue: uncertainty in the audio and fixed harmonic templates from the score

Incremental audio-to-score alignment, also known as *score following*, methods are essential to automatic accompaniment systems [5], intelligent score viewers [2], and robot musicians [13] because the alignment synchronizes these systems with human performances. We need a probabilistic framework for the audio-to-score alignment problem in order to cope with the uncertainty in the audio signal generated from the score in a symbolic representation.

Existing methods tend to fail the alignment when multiple musical notes are played by multiple musical instruments. That is, the audio signal contains various timbres and the volume ratio of each musical note is unsure. Figure 1 illustrates this issue. The observed pitched audio signal includes equally-spaced peaks in frequency domain called a harmonic structure. The observed audio is matched with harmonic structure templates generated from the score. Musical notes written in the score is played with arbitrary musical instruments. The resulting audio harmonic structures can vary from instrument to instrument whereas the templates of the score have been set in advance using some heuristics or a parameter learning [4]. In Figure 1, harmonic structures of a guitar and a violin is shown in blue and red lines, respectively. Furthermore, the mixture ratio of each note in the audio is unknown until the observation while the ratio in the template is fixed, typically equal.

Thus, the variety of the audio signal causes a mismatch

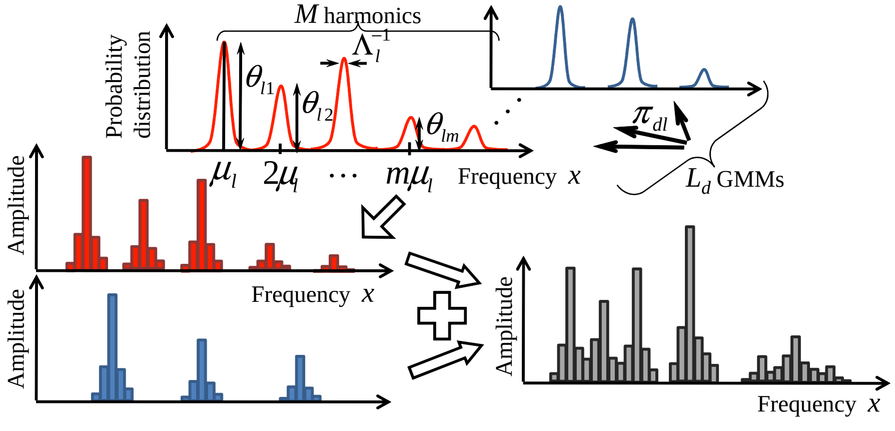

Figure 2. Audio spectrogram based on LHA

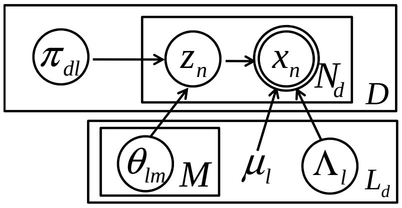

Figure 3. Graphical model of LHA

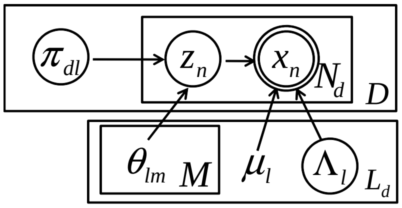

Figure 4. LHA with fixed θ_{lm}'s

between the observed harmonic structure and the fixed one generated from the score. We need a flexible harmonic structure model to robustly match the audio and score since the audio signal is almost unknown until we observe it.

Our idea is to employ a Bayesian harmonic structure model called latent harmonic allocation (LHA) [15]. This model allows us to form harmonic structure templates reflecting the observed audio with the prior knowledge written in the score, e.g., fundamental frequencies of musical notes.

1.1 Related work

Two important aspects reside in modeling audio-to-score alignment: (1) a temporal model of musical notes and (2) an observation model of the input audio signal from the corresponding score. Although improvements are made repeatedly for the temporal model, misalignments are often caused by static and fixed audio observation models. The audio observation model used in the methods introduced in this section uses static features such as chroma vectors or fixed harmonic structure templates based on Gaussian mixture model (GMM). These features are often heuristically designed and therefore lose robustness against uncertain situations in which many instruments are involved and the audio is polyphonic.

Most audio-to-score alignment methods employ dynamic time warping (DTW) [2,6], hidden Markov models (HMM) [4, 12], or particle filters [7, 11, 13]. DTW or HMM-based methods sometimes fails the alignment since the length of musical notes is less constrained in the decoding.

The note length corresponds to the length of a state sequence in the HMM. Cont's method [3] uses a hidden semi-Markov model (HSMM) to control state lengths. The HSMM restricts the duration of a stay at one state so that the state length is limited. While the model refrains from delayed state transitions, this has no restriction on fast transitions. As a result, the HSMM tends to estimate the audio signal faster than it is.

Some methods estimate not only the score position but also the tempo, i.e., the speed of the music for the temporal accuracy. Raphael's method includes the tempo of the music as a state [14] to accurately decode the note lengths. Otsuka et al. [13] propose a particle filter-based method for their simultaneous estimation. While Raphael's method observes only harmonic structures as pitch information, Otsuka et al.'s method observes the periodicity of the onsets to directly estimate the tempo.

2. AUDIO OBSERVATION MODEL

This Section describes how the audio is generated in terms of LHA. We focus on harmonic structures to associate an audio signal with a symbolic score. The LHA model flexibly fits the shape of harmonic structures given an audio signal observation using variational Bayes inference.

The harmonic peaks are often modeled as a Gaussian mixture model (GMM) by regarding each peak as a single Gaussian [3, 13, 14]. The black lines in Figure 1 are the GMM curves. These methods use Kullback-Leibler divergence (KL-div) as a matching function between the audio harmonics and the GMM template harmonics generated from the score by regarding the harmonic structure as a probability distribution. The mean value of each Gaussian peak is determined by a pitch specified in the score.

LHA [15] is a generative model for harmonic structures of pitched sounds. A graphical model for LHA is depicted in Figure 3. In the LHA model, the amplitude of audio harmonics is regarded as a histogram over the frequency bins.

Figures 2 and 3 explain how a mixture of harmonic structures is generated. Variables in a circle are random variables while those without a circle are parameters. Double circled x_n means an observed variable. For each segment d, N_d, frequencies x_n are observed. The audio spectrogram is segmented into d by chords, which are sets of musical notes. To sample each x_n, a L_dM-dimensional multinomial latent variable z_n is sampled as follows. A harmonic structure GMM l

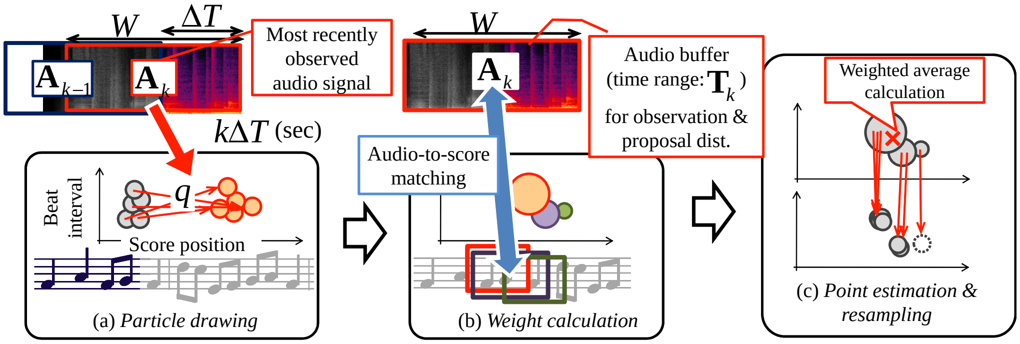

Figure 5. Three steps in particle filtering for the audio-to-score alignment

is selected with probability π_{dl}, where $\sum_{l=1}^{L_d} \pi_{dl} = 1$. Among M Gaussian peaks, m is selected to sample x_n with probability θ_{lm}, where $\sum_{m=1}^{M} \theta_{lm} = 1$. Finally, x_n is sampled from the Gaussian distribution of which mean and precision are $m\mu_l$ and Λ_l, respectively. The definitions of each variable in LHA in Figure 3 are summarized below:

$$p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\Lambda}) = \prod_{dnlm} \mathcal{N}(x_{dn}|m\times\mu_l,\Lambda_l), \tag{1}$$

$$p(\boldsymbol{Z}|\boldsymbol{\pi},\boldsymbol{\theta}) = \prod_{dnlm} (\pi_{dl}\theta_{lm})^{z_{dnlm}}, \tag{2}$$

$$p(\boldsymbol{\pi}) = \prod_{d} \mathrm{Dir}(\boldsymbol{\pi}_d|\boldsymbol{\alpha}_0), \quad p(\boldsymbol{\theta}) = \prod_{l} \mathrm{Dir}(\boldsymbol{\theta}_l|\boldsymbol{\beta}_0), and \tag{3}$$

$$p(\boldsymbol{\Lambda}) = \prod_{l} \mathrm{Gam}(\Lambda_l|a_0,b_0), \tag{4}$$

where $\mathcal{N}(\cdot)$, $\mathrm{Dir}(\cdot)$, $\mathrm{Gam}(\cdot)$ denote the density functions of Gaussian, Dirichlet, and gamma distribution, respectively. The latent variable $z_{dn} = [z_{dnlm}]$ is L_dM-dimensional with one element being 1 and the other being 0. Variables $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ are conjugate priors for $\boldsymbol{Z}$, and the precision of Gaussian harmonics $\boldsymbol{\Lambda}$ is a conjugate prior for $\boldsymbol{X}$. Here, $\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0, a_0$, and b_0 are hyperparameters for each distribution. $\boldsymbol{\alpha}_0 = [\alpha_{0l}]_{l=1}^{L_d}$ is set as $\alpha_{0l} = 1$, and $\boldsymbol{\beta}_0 = [\beta_{0m}]_{m=1}^{M}$ is set as $\beta_{0m} = 1$ because the mixture ratio of each musical note and the height of each harmonic are unknown. A "flat" prior knowledge about these parameters is preferred to reflect our ignorance. The hyperparameters of the gamma distribution are empirically set as $a_0 = 1$ and $b_0 = 2.4$ by considering the width of harmonics determined by the window function of a short-time Fourier transform (STFT).

The LHA is originally designed for multi-pitch analysis [15], and therefore the fundamental frequency μ_l is a random variable. However, in our audio-to-score alignment framework, μ_l is treated as a parameter because fundamental frequencies are given by the score as musical notes. This is why μ_l is not in a circle in Figure 3.

In general, too flexible model can cause an over-fitting problem. LHA is flexible in terms of the mixture ratio $\boldsymbol{\pi}$ and harmonic heights $\boldsymbol{\theta}$. To limit the model complexity, we fix the harmonic heights $\boldsymbol{\theta}$ and only consider the mixture ratio $\boldsymbol{\pi}$ as in Figure 4. We refer to the former model in Figure 3 as *full LHA*, and the latter in Figure 4 as *mixture LHA*.

3. AUDIO-TO-SCORE ALIGNMENT USING PARTICLE FILTER

This section presents the problem setting and procedures of our method. The problem is specified as follows:

Inputs: incremental audio signal and the corresponding whole score
Outputs: the current score position and tempo
Assumptions: (1) The score includes musical notes and the approximate tempos of the music. (2) Musical notes are pairs of their pitch and length, e.g., a quarter note, (3) Approximate tempos are specified as the range of a tempo, e.g., 90–110 beats per minute (bpm).

No prior knowledge about musical instruments is assumed.

3.1 Method overview

Let k be the index of filtering steps and $A_{t,f}$ be the amplitude of the input audio signal in the time-frequency domain. Here, t and f denote the time (sec) and the frequency (Hz), respectively. Our system is implemented at a sampling rate of 44100 (Hz), a window length of 2048 (pt), and a hop size of 441 (pt). $\bar{A}_{t,f}$ denotes a quantized integer amplitude given by $\bar{A}_{t,f} = \lfloor A_{t,f}/\Delta A \rfloor$, where ΔA is the quantization factor, and $\lfloor\cdot\rfloor$ is the flooring function. $\Delta A = 3.0$ in our implementation. This value should be so small that the shape of the spectrum is preserved after the quantization and that sufficient observations are provided for the Bayesian inference in the LHA. Let p (beat) be the score position. The score is divided into frames whose lengths are equal to $1/12$ of one beat, namely, a quarter note[1] . Musical notes are denoted by $\boldsymbol{\mu}_p = [\mu_p^1 ... \mu_p^{L_p}]^T$, where L_p is the number of notes at p, and μ is the fundamental frequency of the note.

Figure 5 illustrates the procedures. At every ΔT (sec), the particle filtering [1] proceeds as: (a) move particles in accordance with elapsed ΔT (sec) by drawing particles from the proposal distribution, (b) calculate the weight of each particle, (c) report the point estimation of the score position and beat interval, and resample the particles. Each particle has

[1] p is discretized at 1/12 interval in (beat).

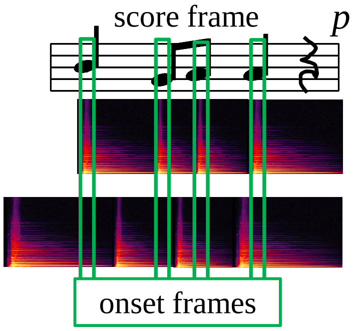

Figure 6. Score position proposal. Audio frames marked by green rectangles are aligned with score onsets.

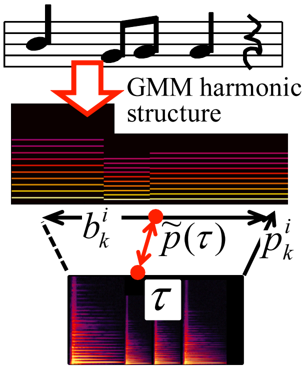

Figure 7. Frame-by-frame alignment of audio using p_k^i and b_k^i

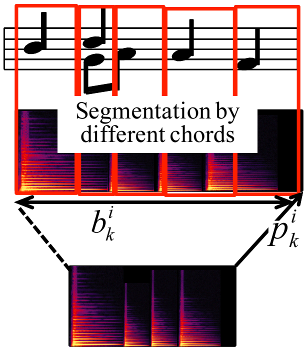

Figure 8. Segmentation by chords for LHA observation

the following information as a hypothesis: the score position p_k^i, beat interval (sec/beat), i.e., the inverse tempo, b_k^i, and the weight w_k^i as a fitness to the model.

In the kth filtering step, the particle filter estimates the posterior distribution of the score position p_k and beat interval b_k given the latest audio spectrogram $\mathbf{A}_k = [A_{\tau,f}]$, where $\tau \in \mathbf{T}_k$, $\mathbf{T}_k = \{t | k\Delta T - W < t \leq k\Delta T\}$, and W is the window length for the audio spectrogram. The posterior distribution is approximated using many particles as in $p(\mathbf{s}_k|\mathbf{A}_k) = \sum_{i=1}^{I} w_k^i \delta(\mathbf{s}_k^i - \mathbf{s}_k)$, where I is the number of particles, and $\mathbf{s}_k^i = [p_k^i, b_k^i]$ denotes the state of the ith particle.[2] The weight of each particle w_k^i is calculated as:

$$w_k^i \quad \propto \quad \frac{p(\mathbf{s}_k^i|\mathbf{s}_{k-1}^i)p(\mathbf{A}_k|\mathbf{s}_k^i)}{q(\mathbf{s}_k^i|\mathbf{s}_{k-1}^i, \mathbf{A}_k)}, \tag{5}$$

where $p(\mathbf{s}_k^i|\mathbf{s}_{k-1}^i)$ and $p(\mathbf{A}_k|\mathbf{s}_k^i)$ in the numerator are the state transition model and observation model, respectively. New score position and beat interval values are drawn at each step from the proposal distribution $q(\mathbf{s}_k^i|\mathbf{s}_{k-1}^i, \mathbf{A}_k)$.

3.2 Drawing particles from the proposal distribution

Particles are drawn from the proposal distribution in Eq. (6). First, a new beat interval b_k^i is drawn, then a new score position p_k^i is drawn depending on the drawn b_k^i. The proposal is designed to draw (1) a beat interval that lies in the tempo range provided by the score and that matches the intervals among audio onsets and (2) a score position that matches the increase of the audio amplitude with the score onset frame.

$$\begin{aligned} \mathbf{s}_k^i \quad &\sim \quad q(b, p|\mathbf{s}_{k-1}^i, \mathbf{A}_k) \\ &\propto \quad R(b; \mathbf{A}_k)\Psi(b; \tilde{b}) \times Q(p; b, \mathbf{A}_k, \mathbf{s}_{k-1}^i). \end{aligned} \tag{6}$$

$R(b; \mathbf{A}_k)$ and $\Psi(b; \tilde{b})$ denote the normalized cross correlation of the audio signal and the window function that limits the range of the beat interval, respectively. $Q(p; b, \mathbf{A}_k, \mathbf{s}_{k-1}^i)$ denotes the onset matching function.Detailed equations are explained in [13].

The onset matching function $Q(p; b, \mathbf{A}_k, \mathbf{s}_{k-1}^i)$ in Eq. (6) represents how well the audio and score are aligned in terms of the onsets. Figure 6 explains the design. The top case in which audio frames with a peak power is aligned with score onsets results in the larger Q, where as the bottom case Q is a small value since the onsets are misaligned. The detailed mathematical expressions are presented in [13].

3.3 Weight calculation

The weight for each particle is calculated with the sampled value $\mathbf{s}_k^i$ in Eq. (5) by using the state transition model,

$$p(\mathbf{s}_k^i|\mathbf{s}_{k-1}^i) \quad = \quad \mathcal{N}(p_k^i|\hat{p}_k^i, \sigma_p^2) \times \mathcal{N}(b_k^i|b_{k-1}^i, \sigma_b^2), \tag{7}$$

and the observation model,

$$p(\mathbf{A}_k|\mathbf{s}_k^i) \quad \propto \quad p(\mathbf{A}_k|p_k^i) \times R(b_k^i; \mathbf{A}_k). \tag{8}$$

The score position transition conforms to a linear Gaussian model with the transition $\hat{p}_k^i = p_{k-1}^i + \Delta T / b_{k-1}^i$ and the variance σ_p^2 (beat^2). The beat interval transition is a random walk model with the variance σ_b^2 (sec^2/beat^2). The variances are empirically set as $\sigma_p^2 = 0.25$ and $\sigma_b^2 = 0.1$, respectively.

For the observation of the beat interval, the normalized cross-correlation of the audio spectrogram, $R(b_k^i; \mathbf{A}_k)$, is again used in Eq. (8). The other factor $p(\mathbf{A}_k|p_k^i)$ is the likelihood corresponding to the pitch information. As explained in Section 2, the GMM-based harmonic structures are used to match the audio and score. First, the matching with KL-div is presented as a baseline where all the GMM parameters, the chord mixture ratio $\boldsymbol{\pi}$ or harmonic heights $\boldsymbol{\theta}$, and the Gaussian width $\boldsymbol{\lambda}$, are fixed. Then, we explain two types of LHA-based audio-to-score matchings, the full LHA and mixture LHA, where the GMM parameters are probability variables that flexibly adapt to the observed audio harmonic structure. Full LHA adapts all $\boldsymbol{\pi}$, $\boldsymbol{\theta}$, and $\boldsymbol{\lambda}$ whereas mixture LHA adapts only $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$ to the audio. Further discussion of the difference is in Section 4.

The KL-div matching uses a normalized amplitude spectrogram $\acute{\mathbf{A}}_k$ while the LHA models use the quantized spectrogram $\bar{\mathbf{A}}_k$. To match the buffered audio, the audio spectrogram $\mathbf{A}_k$ or $\bar{\mathbf{A}}_k$ is aligned with the score shown as Figure 7. As the time $k\Delta T$ is assigned to p_k^i with the beat interval b_k^i, the audio frame τ is linearly assigned to the score frame as given by

$$\tilde{p}(\tau) \quad = \quad p_k^i - (k\Delta T - \tau)/b_k^i. \tag{9}$$

[2] $\delta(\mathbf{x}) = 1$ iff $\mathbf{x} = \mathbf{0}$, otherwise $\delta(\mathbf{x}) = 0$.

3.3.1 Harmonic structure observation based on KL-div

For each score frame p, the GMM template of the harmonic structure is generated from the musical notes $\boldsymbol{\mu}_p$ as:

$$\hat{A}_{p,f} \;=\; \sum_{l=1}^{L_p}\sum_{m=1}^{M} C_{\mathrm{harm}}\pi_l\theta_m\mathcal{N}(f|g\mu_p^l,\sigma_{\mathrm{KL}}^2)+C_{\mathrm{floor}}, \quad (10)$$

where L_p is the number of notes at p and the number of harmonic structures M is 10. The ratio of each note is equally set as $\pi_l = 1/L_p$. The height of the mth harmonic is set as $\theta_m = 0.2^{m-1}$. The variance is set as $\sigma_{\mathrm{KL}}^2 = 2.4$, derived from the window function used in STFT. C_{floor} is a flooring constant to ensure $\hat{A}_{p,f} > 0$ and avoid zero-divides in Eq. (11). $C_{\mathrm{harm}} = 0.95$ and C_{floor} is set such that the harmonic structure template is normalized as $\hat{A}_{p,\cdot} = 1$. The subscript $\cdot$ means a summation over the replaced index.

Here, the audio likelihood using KL-div is defined as

$$\log p(\mathbf{A}_k|\mathbf{s}_k^i) \;=\; -\sum_{\tau\in\mathbf{T}_k}\sum_{f}\acute{A}_{\tau,f}\log\frac{\acute{A}_{\tau,f}}{\hat{A}_{\tilde{p}(\tau),f}}, \quad (11)$$

where $\acute{A}_{\tau,f} = A_{\tau,f}/A_{\tau,\cdot}$ is the normalized amplitude. The right-hand side of Eq. (11) is a negative KL-div between the audio harmonic structure and the GMM harmonic template.

3.3.2 LHA-based likelihood calculation

We first explain how LHA is used as the likelihood, then show the iterations for both full and mixture LHA inferences. The quantized amplitudes $\bar{\mathbf{A}}_k$ are regarded as a histogram of amplitudes over frequency bins $\mathbf{X}$ illustrated as gray bars in Figure 2. The rigorous likelihood in Eq. (5) is

$$p(\mathbf{X}|\mathbf{s}_k^i) \;=\; \sum_{\mathbf{Z}}\iiint p(\boldsymbol{X},Z,\pi,\theta,\Lambda|p_k^i,\boldsymbol{\mu})d\boldsymbol{\pi}d\boldsymbol{\theta}d\boldsymbol{\Lambda}. \quad (12)$$

Since this analytical summation over $\mathbf{Z}$ is intractable [3], we infer the latent variables $\mathbf{Z},\boldsymbol{\pi},\boldsymbol{\theta}$, and $\boldsymbol{\Lambda}$ by variational Bayes (VB) method under the factorization assumption $q(\mathbf{Z},\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\Lambda}) = q(\mathbf{Z})q(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\Lambda})$. We use the variational lower bound for the weight calculation as an approximate observation model instead of Eq. (12),

$$\log p(\mathbf{X}|\mathbf{s}_k^i) \approx \mathcal{L}(q) = \mathbb{E}_{\mathbf{Z},\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\Lambda}}\left[\log\frac{p(\boldsymbol{X},\mathbf{Z},\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\Lambda}|p_k^i,\boldsymbol{\mu})}{q(\mathbf{Z},\pi,\theta,\Lambda)}\right], \quad (13)$$

where $\mathbb{E}_{\mathbf{Z},\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\Lambda}}[\cdot]$ denotes an expectation over $q(\mathbf{Z},\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\Lambda})$. For the inference of LHA, the audio is segmented by the chord in the score as shown in Figure 8. This segmentation d is made on the basis of the alignment by Eq. (9).

The variational lower bound in Eq. (13) is maximized with the following variational posteriors:

$$q(\mathbf{Z}) = \prod_{dnlm}\gamma_{dnlm}^{z_{dnlm}}, \quad q(\boldsymbol{\pi}) = \prod_{d}\mathrm{Dir}(\boldsymbol{\pi}_d|\boldsymbol{\alpha}_d),$$
$$q(\boldsymbol{\theta}) = \prod_{l}\mathrm{Dir}(\boldsymbol{\theta}_l|\boldsymbol{\beta}_l), \quad q(\boldsymbol{\Lambda}) = \prod_{l}\mathrm{Gam}(\Lambda_l|a_l,b_l),$$

the parameters of which are updated as

$$\gamma_{dnlm} = \quad \rho_{dnlm}/\rho_{dn\cdot\cdot}, \quad (14)$$

$$\begin{aligned}\log\rho_{dnlm} =& \psi(\alpha_{dl}) - \psi(\alpha_{d\cdot}) + \psi(\beta_{lm}) - \psi(\beta_{l\cdot}) \\ &+\psi(a_l)/2 - (\log b_l)/2 - (x_n - m\mu_l)^2 a_l/2b_l,\end{aligned} \quad (15)$$

$$\begin{aligned}\alpha_{dl} =& \alpha_{0l} + \gamma_{d\cdot l\cdot}, \quad \beta_{lm} = \beta_{0m} + \gamma_{\cdot\cdot lm}, \\ a_l = a_0 + \frac{\gamma_{\cdot\cdot l\cdot}}{2}&, \quad b_l = b_0 + \frac{\sum_{dnm}\gamma_{dnlm}(x_{dn} - m\mu_l)^2}{2},\end{aligned} \quad (16)$$

where $\psi(\cdot)$ in Eq. (15) denotes the digamma function. Eqs. (14,15) and Eqs. (16) are iteratively calculated until the lower bound in Eq. (13) converges. Note that Eq. (14) is the normalization of ρ over indices l and m.

Mixture LHA update: In the update for the mixture LHA model, the harmonic height parameter is set as $\theta_{lm} = 0.2^{m-1}$. Thus, the update equations are modified as:

$$\begin{aligned}\log\rho_{dnlm} =& \psi(\alpha_{dl}) - \psi(\alpha_{d\cdot}) + \log\theta_{lm} \\ &\psi(a_l)/2 - (\log b_l)/2 - (x_n - m\mu_l)^2 a_l/2b_l,\end{aligned} \quad (17)$$

$$\begin{aligned}\alpha_{dl} =& \alpha_{0l} + \gamma_{d\cdot l\cdot}, \\ a_l = a_0 + \frac{\gamma_{\cdot\cdot l\cdot}}{2}&, \quad b_l = b_0 + \frac{\sum_{dnm}\gamma_{dnlm}(x_{dn} - m\mu_l)^2}{2}.\end{aligned} \quad (18)$$

Relationship with the KL-div likelihood: Remember the negative KL-div is used as the log-likelihood in Eq. (11). The following equation always holds during the iterations:

$$\mathcal{L}(q) + KL(q||p) = \log p(\mathbf{X}|\mathbf{s}_k^i) \quad (\text{const wrt. } q).$$

The KL-div is defined between the approximate distribution $q(\mathbf{Z},\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\Lambda})$ and the true posterior $p(\mathbf{Z},\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\Lambda}|\boldsymbol{X},p_k^i,\boldsymbol{\mu})$. Note that maximizing $\mathcal{L}(q)$ is equivalent to minimizing KL-div, namely, maximizing the negative KL-div due to the equation above. Thus, the LHA-based likelihood is interpreted as an extension of Eq. (11) in that the harmonic templates adapt to the audio observation to minimize the KL-div and maximize the log-likelihood.

3.4 Point estimation and efficient computing

After the weights of all particles are calculated, the point estimation is reported as $\hat{\mathbf{s}}_k = \sum_{i=1}^{I} w_k^i\mathbf{s}_k^i / \sum_{i=1}^{I} w_k^i$. Particles are resampled after the point estimation procedure to eliminate zero-weight particles. The resampling probability is in proportion to the weight of each particle [1].

4. EXPERIMENTAL RESULTS

This section presents the alignment error of three observation models; conventional KL-div [13], full LHA, and mixture LHA. Twenty songs from RWC Jazz music database [8] is used for this experiment. This test set includes various compositions of musical instruments from solo performance to big band ensembles. Our system is implemented on Linux OS and a 2.4 (GHz) processor. Experiments are carried out with the following parameter settings; the filtering interval $\Delta T = 0.5$ (sec), the window length for the audio processing $W = 1.5$ (sec), and the number of particles $I = 300$.

Figure 9 shows the error percentiles of 20 songs for three methods. Black, red, and blue bars represent the percentiles of KL-div, full LHA, and mixture LHA, respectively. The darkest bars are the 50% percentiles, middle bars are the 75%, and the lightest segments are the 100% percentiles. The less values indicate the better performance. Songs with a larger ID tend to involve more instruments. Both of the

[3] The integration over $\boldsymbol{\pi}$, $\boldsymbol{\theta}$, and $\boldsymbol{\Lambda}$ is tractable thanks to their conjugacy.

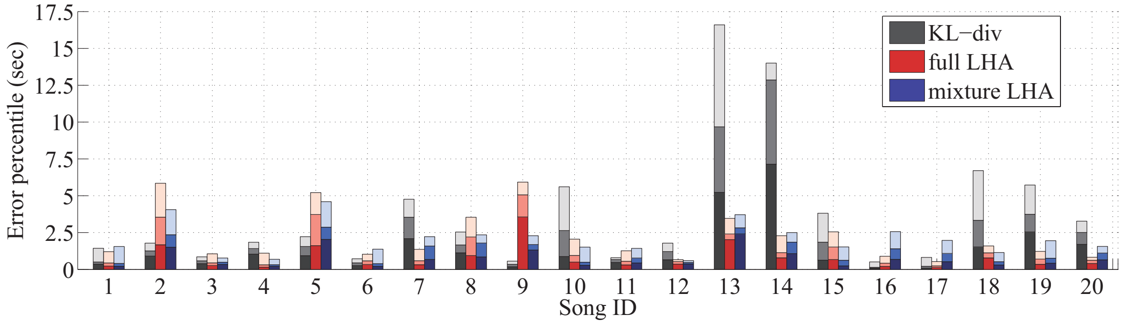

Figure 9. Error percentiles for 20 songs

LHA-based methods outperform the KL-div for 10 songs, and either full or mixture LHA shows less errors than KL-div for 3 songs. In particular, LHA-based methods tend to report less errors when the song consists of a larger number of instruments (larger ID songs). This is what we expect from the LHA models.

Two major reasons are given why LHA-based observation models still accumulate the alignment error. First, LHA is vulnerable to rest notes where no musical note is specified. This is because the LHA model penalizes unspecified harmonic peaks. When the score provides a rest, LHA penalizes any audio observation. The error caused by these rest notes is seen in songs 2, 5, 8, and 9, where we have more chances to have rest notes because the number of musical instruments is relatively small. The second reason is the non-harmonic feature of percussions and drums. Because drum sounds are loud and outstanding in the ensemble, these sounds interfere the harmonic structures of pitched sounds assumed by LHA. This case applies in songs 11, 16, and 17 where drums are included in the ensemble.

Here we discuss the difference between the full and mixture LHAs. Since mixture LHA has less variables to infer, we can expect more accurate inference as long as the fixed parameters $\boldsymbol{\theta}$ fit the observation. The fixed $\boldsymbol{\theta}$ declines as the frequency becomes larger. This descending height is well observed in stringed instruments such as guitar or piano dominantly used in songs 1-6; whereas wind instruments such as saxophone or flute or bowed instruments such as violin show rather different peaks. When these instruments are dominant in a song, e.g., songs 16 and 17 which are in a big band style, the full LHA will be the better choice.

5. CONCLUSION AND FUTURE WORKS

The experiment has shown that LHA is especially effective in a large-ensemble situation where more musical notes are simultaneously performed. However, LHA-based audio observation models is disturbed by (1) rest notes and (2) drum sounds. To make the best use of the LHA model, one promising solution is to examine the musical score in advance of the alignment whether the expecting audio signal is suitable for LHA. The development of this top-level decision making process will be one of the future works.

Another future work includes an accelerated calculation of LHA iterations for such real-time applications as automatic accompaniment systems. Current implementation requires approximately 10 seconds to process one-second audio data.

6. REFERENCES

[1] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking. *IEEE Transactions on Signal Proc.*, 50(2):174–189, 2002.

[2] A. Arzt, G. Widmer, and S. Dixon. Automatic Page Turning for Musicians via Real-Time Machine Listening. In *Proc. of the European Conference on Artificial Intelligence*, pages 241–245, 2008.

[3] A. Cont. A Coupled Duration-Focused Architecture for Realtime Music to Score Alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6):974–987, 2010.

[4] A. Cont, D. Schwarz, and N. Schnell. Training IRCAM's score follower. In *AAAI Fall Symposium on Style and Meaning in Art, Language and Music*, 2004.

[5] R. Dannenberg and C. Raphael. Music Score Alignment and Computer Accompaniment. *Comm. ACM*, 49(8):38–43, 2006.

[6] S. Dixon. An On-line Time Warping Algorithm for Tracking Musical Performances. In *Proc. of the IJCAI*, pages 1727–1728, 2005.

[7] Z. Duan and B. Pardo. A STATE SPACE MODEL FOR ONLINE POLYPHONIC AUDIO-SCORE ALIGNMENT. In *Proc. of Int'l Conf. on Acoustics, Speech and Signal Processing*, pages 197–200, 2011.

[8] M. Goto. AIST Annotation for RWC Music Database. In *Proc. of ISMIR*, pages 359–360, 2006.

[9] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno. Instrument Identification in Polyphonic Music: Feature Weighting to Minimize Influence of Sound Overlaps. *EURASIP Journal on Applied Signal Processing*, vol. 2007, 2007. Article ID 51979.

[10] A. Kulapuri. Multipitch Analysis of Polyphonic Music and Speech Signals Using an Auditory Model. *IEEE Transactions on Audio, Speech and Language Processing*, 16(2):255–266, 2007.

[11] 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 *Proc. of Int'l Conf. on Acoustics, Speech and Signal Processing*, pages 193–196, 2011.

[12] N. Orio, S. Lemouton, and D. Schwarz. Score Following: State of the Art and New Developments. In *Proc. of Int'l Conf. on New Interfaces for Musical Expression*, pages 36–41, 2003.

[13] T. Otsuka, K. Nakadai, T. Takahashi, K. Komatani, T. Ogata, and H. G. Okuno. Real-Time Audio-to-Score Alignment using Particle Filter for Co-player Music Robots. *EURASIP Journal of Advances in Signal Processing*, vol. 2011, 2011. Article ID 384651.

[14] C. Raphael. Aligning music audio with symbolic scores using a hybrid graphical model. *Machine Learning*, 65(2–3):389–409, 2006.

[15] K. Yoshii and M. Goto. Infinite Latent Harmonic Allocation: A Nonparametric Bayesian Approach to Multipich Analysis. In *Proc. of ISMIR*, pages 309–314, 2010.