Predictive Score Following using Particle Filter for Music Robots

*Takuma Otsuka†, Kazuhiro Nakadai‡, Toru Takahashi†, Tetsuya Ogata†, Hiroshi G. Okuno†
† Graduate School of Informatics, Kyoto University
‡ Honda Research Institute Japan Co., Ltd.

Abstract— Our goal is to develop a *co-player* music robot, i.e., a robot that presents a musical expression together with humans. A music interaction requires two important functions: synchronization with the music and musical expression, such as dancing or playing a musical instrument. Many instrument-performing robots are only capable of the latter function, they may have difficulty in playing live with human performers. The synchronization function is critical for the interaction. To enable robots to play a musical instrument or perform a dance in synchronization with the accompanied music, the robot has to predict the coming musical events. This paper presents a predictive score following algorithm that consists of tempo estimation and incremental audio to score alignment using a particle filter. The tempo is estimation by the normalized cross correlation of the audio spectrogram and the audio is aligned with the score using KL-divergence criterion. Experiments are carried out using 20 polyphonic jazz songs performed by human musicians. Our method outperformed an existing score following method for 16 out of 20 songs.

Key Words: Music robot, Score following, Particle filter, Human robot interaction

1. Introduction

Our goal is to develop co-player music robots capable of performing a music together with human musicians to provide richer musical experiences. Their music interaction requires two important functions; synchronization with the music and generation of musical expressions, such as dancing or playing a musical instrument. Many instrument-performing robots such as those presented in [1] are only capable of the latter function, they may have difficulty in playing live with human performers. The former function is essential to the interaction.

Since most of the existing music robots tracks the rhythmic structures in the audio signal, their musical expressions are limited to repetitive or random expressions [2,3]. Although the thereminist robot developed by Mizumoto et al. [4, 5] plays a melodious phrase based on a beat tracking algorithm [2], the robot often misses a musical beat and plays asynchronously with the human drummer.

This paper presents a score following algorithm, an incremental audio-score alignment, that enables robots to track the melody in addition to the rhythm in the music given the corresponding musical score. By this method, the robots are able to directly estimate the position of the music that is currently played and to play their phrases accordingly.

2. Requirements in Score Following for Music Robots

Music robots have to not only *follow* the music but also *predict* coming musical notes. This is because a music robot cannot present a musical expression without any delay. For example, Murata *et al.* [2] reports that it takes around 200 (ms) to generate a singing voice using singing voice synthesizer VOCALOID. The thereminist robot [4] also requires around 300 – 500 (ms) to move its arm to play musical notes. Therefore, a robot for our purpose needs the capability to predict future musical events at least 200 – 500 (ms) in advance.

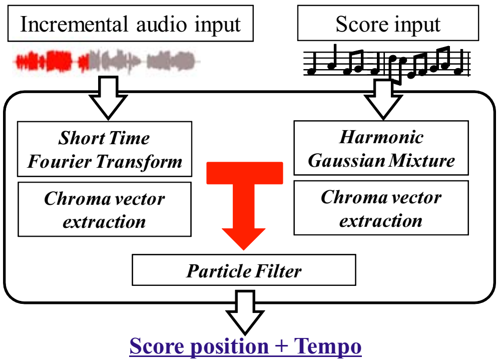

Fig.1: Two-level synchronization architecture

Recently proposed score following method [6] is capable of prediction by estimating both the tempo and score position of the music. This tempo estimation assumes that the alignment between the audio and score is successful. Therefore, the prediction is often erroneous in case the music is polyphonic and that the alignment is difficult. Our method is an extension of the particle filter-based score following [7] that is apt to fail when the tempo is misestimated. We use a prior tempo information specified by the score to stabilize the tempo estimation.

Problem Statement

Input: incremental audio signal and the corresponding musical score,
Output: the score position & tempo,
Assumption: The tempo is provided by the musical score with a margin of error.

Generally, the tempo given by the score and the actual tempo in the human performance is different due to the preference of the performer or the interpretation of the song. Therefore, some margin of error should be assumed in the tempo information. In Section 4·2 several values of the margin are tested.

We model this simultaneous estimation as a state-space model using a particle filter. Figure 1 outlines our method. The particle filter outputs two types of information: the score position and tempo. The future score position is predicted by extrapolating the score position with the current tempo.

3. Our algorithm

Let $X_{f,t}$ be the amplitude of the input audio signal in the time frequency domain with frequency bin f and time t, and let k be the score frame. The score is divided into frames such that the length of a quarter note equals to 12 frames to account for the resolution of sixteenth-note and triplets. Musical notes $\mathbf{n}_k = [n_k^1 ... n_k^{r_k}]^T$ are placed at frame k, and r_k is the number of musical notes. Each particle p_i has score position, beat interval, and weight: $p_i = (\hat{k}_i, \hat{b}_i, w_i)$, and N is the number of particles, i.e., $1 \leq i \leq N$. The units are $\hat{k}_i$ (beat) and $\hat{b}_i$ (sec/beat).

At every ΔT time, the following procedure is carried out as illustrated in Figure 2: (1) state transition, (2) observation, (3) resampling, and then estimation of the tempo and score position. The particle size represents its weight. After the resampling step, the weights of all particles are set to be equal.

3·1 State Transition Model

The beat interval is sampled from the proposal distribution $q(b|\mathbf{X}_t, \tilde{b}^s)$ that consists of normalized cross correlation (NCC) of an audio spectrogram $\mathbf{X}_t$ and the window function derived from the tempo $\tilde{b}^s$ provided by the musical score.

$$\hat{b}_i \sim q(b|\mathbf{X}_t, \tilde{b}^s), \qquad (1)$$

$$q(b|\mathbf{X}_t, \tilde{b}^s) \propto R(b, \mathbf{X}_t) \times \psi(b|\tilde{b}^s). \qquad (2)$$

The audio spectrogram is denoted by $\mathbf{X}_t = [X_{f,\tau}]$, where $t - L < \tau \leq t$ and L denotes the window length of the spectrogram. The NCC is defined as

$$R(b, \mathbf{X}_t) = \frac{\sum_{\tau=t-L}^{t} \sum_{f} X_{f,\tau} X_{f,\tau-b}}{\sqrt{\sum_{\tau=t-L}^{t} \sum_{f} X_{f,\tau}^2 \sum_{\tau=t-L}^{t} \sum_{f} X_{f,\tau-b}^2}}. \qquad (3)$$

The window function is centered at $\tilde{b}^s$ that is the tempo specified by the musical score.

$$\psi(b|\tilde{b}^s) = \begin{cases} 1 & |60/b - 60/\tilde{b}^s| < W \\ 0 & \text{otherwise} \end{cases}, \qquad (4)$$

where W is the width of the window in beats per minute (bpm). A beat interval b (sec/beat) is converted into a tempo value m (bpm=beat/min) by the equation $m = 60/b$. Eq. (4) limits the beat interval value of particles so as not to miss the score position by a falut tempo estimation. The score position is sampled from the normal distribution whose mean value is obtained by adding an offset corresponding to the beat interval $\hat{b}_i$ to the previous score position.

$$\hat{k}_i \sim \mathcal{N}(k|\hat{k}_i^{old} + \Delta T/\hat{b}_i, \sigma_k^2), \qquad (5)$$

where $\hat{k}_i^{old}$ is the previous score position, and the variance σ_k^2 is empirically set to 1.

State transition probabilities are defined as follows:

$$\begin{aligned} & p(\hat{b}_i, \hat{k}_i|\hat{b}_i^{old}, \hat{k}_i^{old}) \\ & = \mathcal{N}(\hat{b}_i|\hat{b}_i^{old}, \sigma_b^2) \times \mathcal{N}(\hat{k}_i|\hat{k}_i^{old} + \Delta T/\hat{b}_i, \sigma_k^2), \end{aligned} \qquad (6)$$

where the variance for the beat interval transition σ_b^2 is empirically set to 0.2. These probabilities are used for the weight calculation in Eq. (7).

3·2 Observation Model

At time t, a spectrogram $\mathbf{X}_t = [X_{f,\tau}](t - L < \tau \leq t)$ is used for the weight calculation. The weight of each particle $w_i, 1 \leq i \leq N$ is calculated as

$$w_i = \frac{p(\mathbf{X}_t|\hat{b}_i, \hat{k}_i) p(\hat{b}_i, \hat{k}_i|\hat{b}_i^{old} \hat{k}_i^{old})}{q(b|\mathbf{X}_t, \tilde{b}^s)}. \qquad (7)$$

The observation probability $p(\mathbf{X}_t|\hat{b}_i, \hat{k}_i)$ consists of three parts as

$$p(\mathbf{X}_t|\hat{b}_i, \hat{k}_i) \propto w_i^{ch} \times w_i^{sp} \times w_i^t. \qquad (8)$$

The two weights, the chroma vector weight w_i^{ch} and spectrogram weight w_i^{sp}, are measures of pitch information. The weight w_i^t is a measure of temporal information.

To match the spectrogram $X_{f,\tau}$, where $t - L < \tau \leq t$, the audio sequence is aligned with the corresponding score for each particle, as shown in Figure 3. Each frame of the spectrogram at time τ is assigned to the score frame k_τ^i that is discrete at 1/12 interval using the estimated score position $\hat{k}_i$ and the beat interval (tempo) $\hat{b}_i$ as:

$$k_\tau^i = \frac{1}{12} \lfloor 12 \times (\hat{k}_i - (t - \tau)/\hat{b}_i) + 0.5 \rfloor, \qquad (9)$$

where $\lfloor x \rfloor$ is the floor function.

A chroma vector has 12 elements corresponding to the pitch name, $C, C\sharp, ..., B$. The sequence of chroma

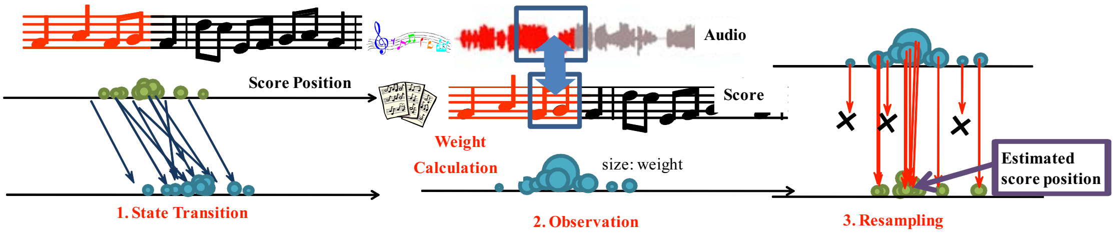

Fig.2: Overview of the Score Following using Particle Filter

vectors $\mathbf{c}^a_\tau$ is calculated from the spectrum $X_{f,\tau}$ using 12 types of band-pass filters for each element [8]. The value of each element in the score chroma vector $\mathbf{c}^s_{k^i_\tau}$ is 1 when the score has a corresponding note, and 0 otherwise. The chroma weight w^{ch}_i is calculated as:

$$w^{ch}_i = \frac{1}{L_{frm}} \sum_{\tau=t-L}^{t} \mathbf{c}^a_\tau \cdot \mathbf{c}^s_{k^i_\tau}, \tag{10}$$

where L_{frm} is the number of audio frames equivalent to L (sec). Both vectors $\mathbf{c}^a_\tau$ and $\mathbf{c}^s_{k^i_\tau}$ are normalized before applying them to Eq. (10).

The spectrogram weight w^{sp}_i is derived from the Kullback-Leibler divergence with regard to the shape of spectrum between the audio and the score.

$$w^{sp}_i = \left(1 + D^{KL}_i\right) \exp\left(-D^{KL}_i\right), \tag{11}$$

$$D^{KL}_i = \frac{1}{L_{frm}} \sum_{\tau=t-L}^{t} \sum_f X_{f,\tau} \log \frac{X_{f,\tau}}{\hat{X}_{f,k_{\tau i}}}, \tag{12}$$

where D^{KL}_i in Eq. (12) is the dissimilarity between the audio and score spectrograms. Before calculating Eq. (12), the spectrum is normalized such that $\sum_f X_{f,\tau} = \sum_f \hat{X}_{f,k^i_\tau} = 1$. The positive value D^{KL}_i is mapped to the weight w^{sp}_i by Eq. (12) where the range of w^{sp}_i is between 0 and 1. For the calculation of w^{sp}_i, the spectrum $\hat{X}_{f,k^i_\tau}$ is generated from the musical score by using the harmonic gaussian mixture model (GMM), the first term in Eq. (13).

$$\hat{X}_{f,k^i_\tau} = \sum_{r=1}^{r_{k_\tau i}} \sum_{g=1}^{G} h(g) N(f; gF_{n^r_{k_\tau i}}, \sigma^2) + C(f), \tag{13}$$

$$C(f) = A \exp\left(-\alpha f\right). \tag{14}$$

$F_{n^r_{k_\tau i}}$ is the fundamental frequency of note $n^r_{k_\tau i}$ and the variance σ^2. The parameters are empirically set as: $G = 10$, $h(g) = 0.2^g$, $\sigma^2 = 0.8$. Eq. (14) is added to avoid zero divides in Eq. (12). A and α is empirically set 0.013 and 0.024, respectively.

The weight w^t_i is the measure of the beat interval and obtained from the NCC of the spectrogram through a shift by $\hat{b}_i$:

$$w^t_i = R(\hat{b}_i, \mathbf{X}_t), \tag{15}$$

where $R(\hat{b}_i, \mathbf{X}_t)$ is defined in Eq. (3).

3·3 Resampling and Estimation

After calculating the weight of all particles, the particles are resampled. A particle p is drawn independently N times from the distribution:

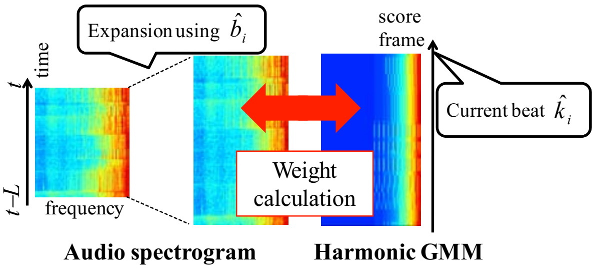

Fig.3: Weight calculation for pitch information

$$P(p = p_i) = \frac{w_i}{\sum_{i=1}^{N} w_i}. \tag{16}$$

After N particles are resampled, the beat interval, equivalent to the tempo, $\hat{b}$ and the score position $\hat{k}$ are estimated by averaging the values that densely distributed particles hold. Then, the score position ΔT ahead in time $\hat{k}^{pred}$is predicted by the following equation:

$$\hat{k}^{pred} = \hat{k} + \Delta T / \hat{b}. \tag{17}$$

3·4 Initial Probability Distribution

The initial particles are set as follows: (1) Draw N samples of the beat interval $\hat{b}_i$ value from a uniform distribution ranging from $\tilde{b}^s - 60/W$ to $\tilde{b}^s + 60/W$ where W is the window width in Eq. (4). (2) Set the score position of each particle $\hat{k}_i$ to 0.

4. Experimental Evaluation

This section presents the prediction error of the score following in various conditions: (a) comparisons with Antescofo [6] and (b) the effect of the width of window function W in Eq. (4).

4·1 Experimental Setup

We used 20 jazz songs from the RWC Music Database [9]. The sampling rate was 44100 (Hz) and Fourier transform was executed with a 2048 (pt) window length and 441 (pt) window shift. The parameter settings are listed in Table 1.

Table 1: Parameter settings

Denotation		Value	
Look-ahead time	ΔT	1	(sec)
Window length	L	2.5	(sec)
Score position variance	σ^2_k	1	(beat^2)
Beat duration variance	σ^2_b	0.2	(sec^2/beat^2)
The number of particles	N	500	(ϕ)

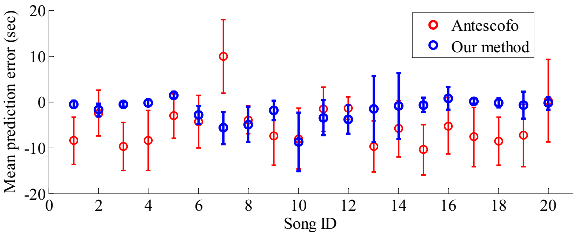

Fig.4: Mean prediction errors in our method and Antescofo, $N = 500, W = 15$ (bpm)

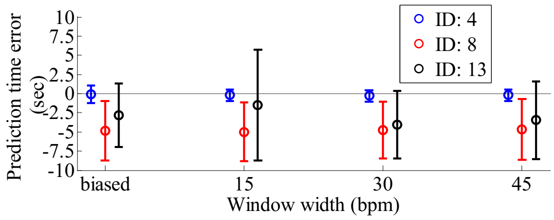

Fig.5: Window width W vs prediction errors

4·2 Score Following Error

At ΔT intervals, our system predicts the score position $\hat{k}(t + \Delta T)$ when the current time is t. Let $s(k)$ be the ground truth time at beat k in the music. $s(k)$ is defined for positive continuous k by linear interpolation of musical event times. The prediction error $e^{pred}(t)$ is defined as:

$$e^{pred}(t) = t + \Delta T - s(\hat{k}(t + \Delta T)). \qquad (18)$$

Positive $e^{pred}(t)$ indicates the estimated score position is ahead of the true position.

Our method vs Antescofo Figure 4 shows the average errors throughout each song in the predicted score positions for 20 songs with $N = 500$ and $W = 15$ (bpm). The comparison between our method in blue plots and Antescofo [6] in red plots. Our method reports less mean error values for 16 our of 20 songs than existing score following algorithm Antescofo.

There can be observed striking errors in songs ID 6–14. Main reasons are two-fold: (1) In songs ID 6–10, a guitar or multiple instruments are used. This multipitch feature makes the audio-score matching difficult. (2) On top of the first reason, temporal fluctuation is observed in songs ID 11–14.

Prediction error vs the tempo width Figure 5 shows the mean prediction errors for various widths of tempo window W. In this experiment, W is set to 15, 30, and 45 (bpm). To simulate the situation that the given tempo is different from the performance, a *biased* case is also tested. In this case, W is set to 15 and the tempo given by the score is biased by 15 (bpm). Therefore, the true tempo is located at the edge of the window function is Eq. (4). Intuitively, the narrower the width is, the closer to zero the error value should be because the chance of choosing a wrong tempo will be reduced. However, almost the same results are obtained for various W and even in the biased case. This is because peaks in the normalized cross correlation in Eq. (3) are sufficiently striking to choose an appropriate beat interval value from the proposal distribution in Eq. (2).

Although the effect of the particle number is tested, the results are almost the same with $N = 300, 500$, or 1000. This indicates the prediction errors are caused by other reasons such as mismatch between the audio and the score in the observation step.

5. Conclusion

This paper presented a particle filter-based score following to attain the synchronization in the musical melody for interactive music robots that presents musical expressions. The experimental results confirmed that our method outperforms an existing score following method and demonstrated the feasibility of the system. Future works include (1) an integration with visual cues that human performer often provides and (2) a quick recovery by searching for a landmark in the music in case of losing the score position.

ACKNOWLEDGMENT: This research was supported by JSPS Grant-in-Aid for Scientific Research in part (S) 19100003 and in part on Innovative Areas (No. 22118502) from the MEXT, Japan.

References

[1] A. Alford *et al.* A music playing robot. In *FSR 99*, pp. 29–31, 1999.

[2] K. Murata *et al.* A Robot Uses Its Own Microphone to Synchronize Its Steps to Musical Beats While Scatting and Singing. In *Proc. of IROS*, pp. 2459–2464, 2008.

[3] G. Weinberg *et al.* Toward Robotic Musicianship. *Computer Music Journal*, Vol. 30, No. 4, pp. 28–45, 2006.

[4] T. Mizumoto *et al.* Human-Robot Ensemble between Robot Thereminst and Human Percussionist using Coupled Oscillator Model. In *Proc. of IROS*, 2010. *to appear.*

[5] T. Otsuka *et al.* Music-ensemble robot that is capable of playing the theremin while listening to the accompanied music. *Trends in Applied Intelligent Systems*, Vol. LNAI 6096, pp. 102–112, 2010.

[6] A. Cont. A Coupled Duration-Focused Architecture for Realtime Music to Score Alignment. *IEEE Trans. on PAMI*, Vol. 32, , 2010. to appear.

[7] T. Otsuka *et al.* Design and Implementation of Two-level Synchronization for Interactive Music Robot. In *Proc. of AAAI*, pp. 1238–1244, 2010.

[8] M. Goto. A Chorus Section Detection Method for Musical Audio Signals and Its Application to a Music Listening Station. *IEEE Trans. on ASLP*, Vol. 14, No. 5, pp. 1783–1794, 2006.

[9] M. Goto *et al.* RWC Music Database: Music Genre Database and Musical Instrument Sound Database. In *Proc. of ISMIR*, pp. 229–230, 2003.