AUTOMATIC MUSIC ACCOMPANIMENT BASED ON AUDIO-VISUAL SCORE FOLLOWING

Akira Maezawa, Kazuhiko Yamamoto
Yamaha Corporation

ABSTRACT

We present an automatic accompaniment system for synchronizing a player piano to human performers, by recognizing both visual and auditory cues. To this end, we extend score following technique to incorporate audio and visual cue detectors. The system generates player piano output and visual cues based on the predicted playback position. The playback position is obtained by coupling the predicted position of human player and a generative model of the accompaniment's tempo track. We present a video demonstration of the system used in a real-life concert.

1. INTRODUCTION

Automatic accompaniment is a method to generate accompaniment to a known piece of music, in such a way that it is synchronized to human musicians. This is achieved by modeling how human musicians coordinate their timings in a music ensemble. Typically, a musician would use auditory and visual cues to track other members in the ensemble. Then, he/she would play the assigned part such that it is synchronized to the ensemble, so that the played part sounds natural. When necessary, the musician would provide physical gestures.

We present a system that mimics these capabilities. To recognize the auditory and visual cues, we extend score following to incorporate not only audio but also visual cue detection. To generate auditory and visual response, we generate accompaniment using player piano and cue gestures that are displayed on screen visible from the players.

2. METHOD

Our system comprises of multimodal score following module, ensemble timing coordination module, and output generation module.

2.1 Score following

The score following is based on segmental HMM (sHMM) that jointly tracks the tempo and the score position. Specifically, we treat the music score as a sequence of R segments. The sHMM expresses the segment r, the expected duration n (in frames) that is used to play the segment, and the current frame l within each segment. The transition of state variables (r, n, l) is then expressed as follows:

1. Self-transition: p
2. Transition from $(r, n, l < n)$ to $(r, n, l + 1)$: $1 - p$
3. Transition from $(r, n, n - 1)$ to $(r + 1, n', 0)$: $(1 - p)\frac{1}{2\lambda^{(T)}} e^{-\lambda^{(T)}|n'-n|}$.

This kind of model is capable of expressing the smoothness of tempo curve, while allowing the duration of each segment to deviate from the expected duration n. In this respect, it can be thought of as using the best of the worlds of an explicit-duration HMM [1] and a left-to-right HMM approach [5]. Each state (r, n, l) has a corresponding position of the score, denoted $\tilde{s}(r, n, l)$.

Based on these parameters, let us express the likelihood of the auditory cues. For each position s of the score, we define the expected normalized constant-Q transform (CQT), $\bar{\mathbf{c}}_\mathbf{s} \in \mathbb{R}^F$, and its normalized half-wave rectified first-order difference, $\mathbf{\Delta}\bar{\mathbf{c}}_\mathbf{s} \in \mathbb{R}^F$. We also define the respective inverse variances $\kappa_s^{(c)}$ and $\kappa_s^{(\Delta c)}$. Then, the likelihood of observing the normalized CQT $\mathbf{c} \in \mathbb{R}^F$ and the normalized ΔCQT $\Delta\mathbf{c} \in \mathbb{R}^F$ is computed as follows:

$$p(c, \Delta c | s = \tilde{s}(r, n, l), \lambda, \{\bar{c}_s\}_{s=1}^S, \{\Delta\bar{c}_s\}_{s=1}^S) = \text{vMF}(c|\bar{c}_s, \kappa_s^{(c)})\text{vMF}(\Delta c|\Delta\bar{c}_s, \kappa_s^{(\Delta c)}). \quad (1)$$

Here, vMF$(\cdot)$ is the von Mises-Fisher distribution.

$\bar{c}$ is generated by analyzing the score data and expressing the expected CQT as a weighted sum of spectral bases; $\Delta\bar{c}$ is generated by taking the normalized half-wave rectified first order difference of $\bar{c}$. Using audio data alone, the model achieves piecewise precision of 96% for chamber music in RWC classical music database [3].

Next, let us discuss the computation of visual cues. Since visual information is important when auditory cues are unavailable [2], detected visual cues are integrated into the observation likelihood. To detect visual cues, we place cameras in front of human musicians and look for a large vertical movement by analyzing the optical flow. An upward motion that is above a given threshold is detected as the start of a cue. To incorporate the cue, the score is marked with annotations $\{\hat{q}_i\}$ that indicate where visual cues are expected. When a cue is detected, we set the likelihood of score positions $\cup[\hat{q}_i - \tau, \hat{q}_i]$ to zero, which

prevents the tracked result to be at slightly before the annotated cue positions.

To compute the posterior distribution to the sHMM in real-time, delayed-decision forward-backward algorithm is used. Then, the mode of the posterior distribution is then approximated as a Normal distribution with mean μ_t and variance σ_t^2.

2.2 Ensemble timing coordination

In order for the accompaniment to respond resonably, it is important to both consider the generative model of human players and the machine, similar in spirit to [6].

We treat the trajectory of the score position played by humans and the machines as a linear dynamical system, assuming that the score position x changes with a slowly drifting velocity v. We assume such dynamical system exists independently for (1) the parts played by the human performers and (2) the machine. Then, the systems are coupled together to produce the final timing. When the nth onset of either the machine or the human players is detected, the system updates its interval variables.

The trajectory of the score position for human performers is modeled as follows:

$$\begin{pmatrix} v_n^{(h)} \\ x_n^{(h)} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \Delta T & 1 \end{pmatrix} \begin{pmatrix} v_{n-1}^{(h)} \\ x_{n-1}^{(h)} \end{pmatrix} + \epsilon_n^{(h)}, \qquad (2)$$

where ΔT is the time elapsed between $n-1$th onset and nth onset. The noise is drawn from a correlated Gaussian noise, since the change in tempo correlates with change in expected score position. Then, the system assumes that at the corresponding onset time, observation μ_t (obtained from score following) is generated from distribution $\mathcal{N}(x_n^{(h)}, \sigma_t^2)$.

The trajectory of the score position for the machine part is given as follows:

$$\begin{pmatrix} v_n^{(a)} \\ x_n^{(a)} \end{pmatrix} = \begin{pmatrix} 1-\beta & 0 \\ \Delta T & 1 \end{pmatrix} \begin{pmatrix} v_{n-1}^{(a)} \\ x_{n-1}^{(a)} \end{pmatrix} + \begin{pmatrix} \beta \bar{v}_n^{(a)} \\ 0 \end{pmatrix} + \epsilon_n^{(a)}. \quad (3)$$

The parameter β controls how strongly the accompaniment part "wants" to play with tempo $\bar{v}_n^{(a)}$. We have previously shown that, in the context of multiple audio alignment, this kind of mean-reverting dynamics is effective at modeling the variability of tempo [4].

Finally, the timing is synchronized by coupling $[v_n^{(h)} x_n^{(h)}]$ and $[v_n^{(a)} x_n^{(a)}]$, yielding in the final output $[v_n, x_n]$:

$$\begin{pmatrix} v_n \\ x_n \end{pmatrix} = \gamma \begin{pmatrix} v_n^{(a)} \\ x_n^{(a)} \end{pmatrix} + (1-\gamma) \begin{pmatrix} v_n^{(h)} \\ x_n^{(h)} \end{pmatrix}. \qquad (4)$$

γ is adjusted according to the extent of leadership assumed by the human (as opposed to the machine) at a given point in the music.

2.3 Output generation

Based on the inferred dynamical system state variable v_n and x_n, the system predicts the current position to play, taking into account the I/O latency. Based on the predicted position, the system sends necessary messages to the player piano. Furthermore, the accompaniment part contains markers that indicate when the system should provide a visual cue to the performers. When the sequencer encounters this marker, a visualization system generates a nodding-like animation, which the human performers see to coordinate their playing. This kind of cue is relevant when the machine part assumes leadership and wants to express its timing fluctuations (e.g., an agogic accent).

3. DEMONSTRATION

We present a video demonstration of our method used in a real-life concert setting, which took place on 5/19/2016 in the Sogakudo Hall at the Tokyo University of the Arts. In the demonstration, the Scharoun Ensemble of the Berliner Philharmoniker and Yamaha's Disklavier player piano plays Schubert's "Trout" quintet together.

Acknowledgment The concert was supported by the Center of Innovation program from the Japan Science and Technology Agency and the Tokyo University of the Arts.

4. REFERENCES

[1] Arshia Cont, José Echeveste, Jean-Louis Giavitto, and Florent Jacquemard. Correct automatic accompaniment despite machine listening or human errors in Antescofo. In *Proc. ICMC*, 2012.

[2] Werner Goebl and Caroline Palmer. Synchronization of timing and motion among performing musicians. *Music Perception: An Interdisciplinary Journal*, 26(5):427–438, 2009.

[3] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. RWC music database: Popular, classical, and jazz music databases. In *Proc. ISMIR*, pages 287–288, 2002.

[4] Akira Maezawa, Katsutoshi Itoyama, Kazuyoshi Yoshii, and Hiroshi G. Okuno. Unified inter- and intra-recording duration model for multiple music audio alignment. In *Proc. WASPAA*, pages 1–5, 2015.

[5] Riccardo Miotto, Nicola Montecchio, and Nicola Orio. Statistical music modeling aimed at identification and alignment. In *Proc. AdMIRe*, pages 187–212, 2010.

[6] Christopher Raphael. A Bayesian network for real-time musical accompaniment. In *Proc. NIPS*, pages 1433–1439, 2001.