A State Space Model for Online Polyphonic Audio-Score Alignment

Zhiyao DUAN & Bryan PARDO

Northwestern University, EECS Department, Evanston, IL, USA.

zhiyaoduan00@gmail.com • pardo@northwestern.edu • http://music.cs.northwestern.edu

Introduction

▶ An **online** algorithm for **polyphonic**, **multi-instrumental** audio-score alignment

▶ Views audio performances as paths in a **2-d space** (score position, tempo)

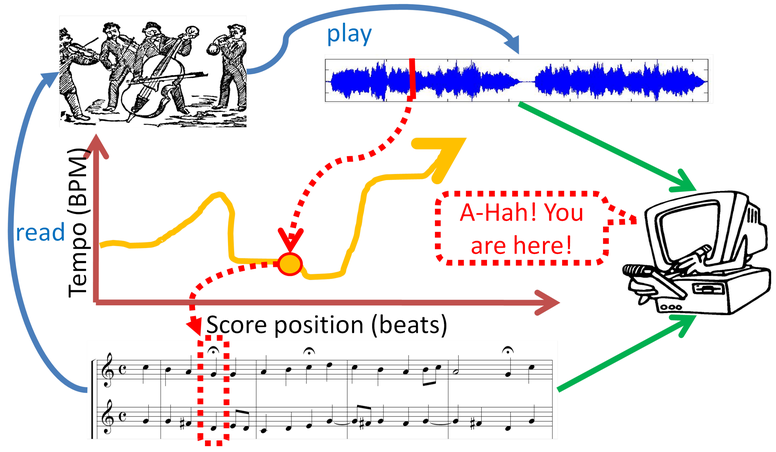
▶ Uses a hidden Markov process model, infered by particle filtering

▶ Performs well on 150 real music pieces

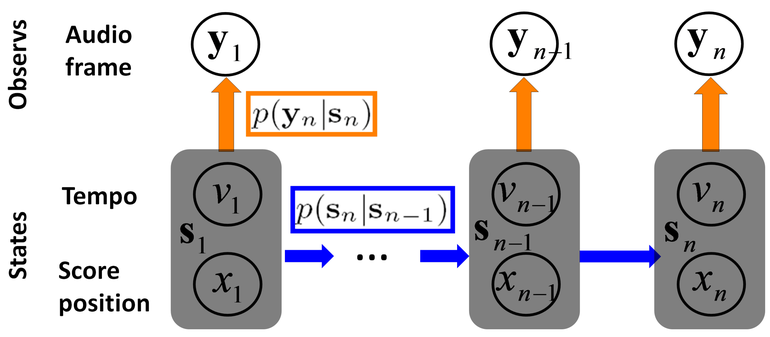
▶ Opens potential applications to **real-time** audio performance editing, source separation, etc.

Problem Formulation

▶ View Audio Performances as Paths



Hidden Markov Process Model



- Aim: estimate $p(\mathbf{s}_n|\mathbf{Y}_{1:n})$ and output $E(x_n)$ for each n, where $\mathbf{Y}_{1:n} = (\mathbf{y}_1, \cdots, \mathbf{y}_n)$

▶ Transition Model: $p(\mathbf{s}_n|\mathbf{s}_{n-1})$

$$x_n = x_{n-1} + l \cdot v_{n-1} \quad (1)$$

$$v_n = \begin{cases} v_{n-1} + n_v & \text{if } x_n \text{ just passes a score note onset/offset} \\ v_{n-1} & \text{otherwise} \end{cases} \quad (2)$$

- l: audio frame hop size (46ms), $n_v \sim \mathcal{N}(0, (\frac{1}{4} \text{ score tempo})^2)$
- No noise term added in Eq.(1) ⇒ Smooth score position estimation
- n_v only introduced in the first line of Eq.(2), ∵ tempo changes can only be perceived near an onset or offset

▶ Observation Model 1: Multi-pitch-based $p(\mathbf{y}_n|\mathbf{s}_n)$

- Find out the set of score pitches $\boldsymbol{\theta}_n$ at position x_n
- Generative: score position x_n ⇒ score pitches $\boldsymbol{\theta}_n$ ⇒ audio frame $\mathbf{y}_n$
- Calculate $p(\mathbf{y}_n|\boldsymbol{\theta}_n)$, i.e. multi-pitch likelihood defined in [1]. Define

$$p(\mathbf{y}_n|\mathbf{s}_n) = -\frac{C}{\ln(p(\mathbf{y}_n|\boldsymbol{\theta}_n))} \quad (3)$$

- Do not make $p(\mathbf{y}_n|\mathbf{s}_n) = p(\mathbf{y}_n|\boldsymbol{\theta}_n)$ ⇒ balance effects of transition model and observation model

▶ Observation Model 2: Chroma-based $p(\mathbf{y}_n|\mathbf{s}_n)$

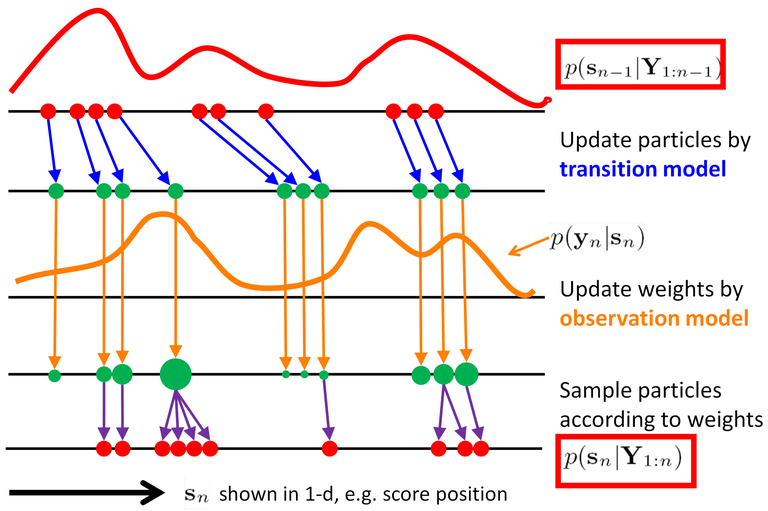
- Represent each audio frame with a 12-d chroma vector $\mathbf{c}_a$
- Calculate for each score position a 12-d chroma vector $\mathbf{c}_m$
- Generative: score position x_n ⇒ score chroma $\mathbf{c}_m$ ⇒ audio chroma $\mathbf{c}_a$

$$\alpha = \arccos\left(\frac{\mathbf{c}_a^T \mathbf{c}_m}{\|\mathbf{c}_a\|\|\mathbf{c}_m\|}\right) \quad (4)$$

$$p(\mathbf{y}_n|\mathbf{s}_n) = C' p(\alpha) = C' \frac{1}{\sqrt{2\pi}} \exp\{-\alpha^2\} \quad (5)$$

- Cosine angle distance instead of Euclidean ⇒ loudness insensitive

▶ Inference: Particle Filtering with 1000 particles



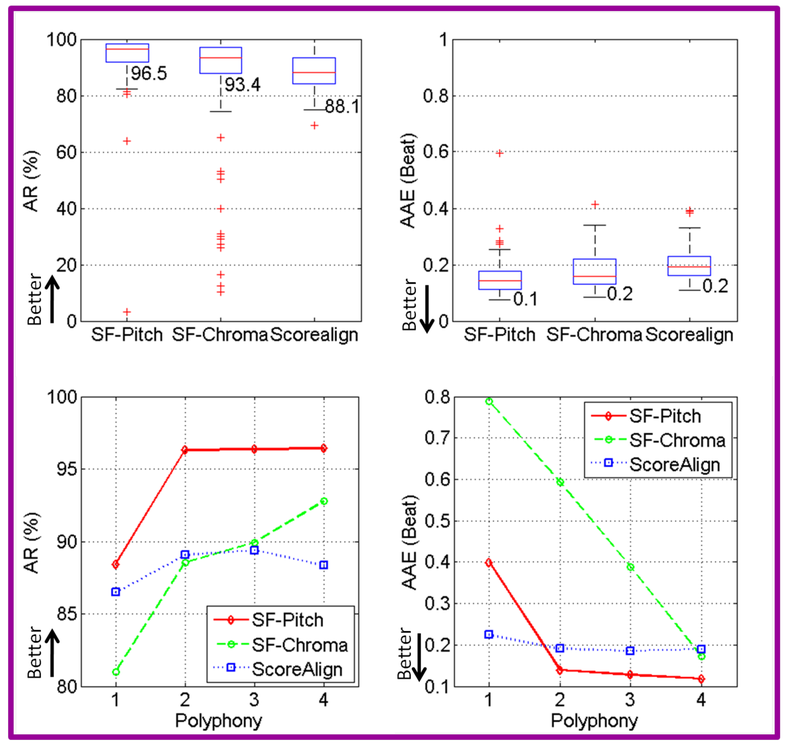
Experiments

▶ Dataset and Measures

- 150 pieces (40 solos, 60 duets, 40 trios, 10 quartet), generated from 10 J.S. Bach 4-part chorales (violin, clarinet, saxophone, bassoon)
- Performance contains fermata; tempo varies within a piece
- Ground-truth alignment obtained manually
- Align Rate (AR): percentage of correctly aligned notes (< 250ms)
- Average Alignment Error (AAE): average absolute alignment deviation of audio frames (unit: beat)

▶ Comparison Methods

- SF-Pitch: proposed using observation model 1
- SF-Chroma: proposed using observation model 2
- ScoreAlign: an open source offline system based on [2]



▶ **Examples:** `http://www.cs.northwestern.edu/~zdu459/icassp2011/examples`

[1] Z. Duan, B. Pardo and C. Zhang, "Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions," *IEEE T-ASLP*, vol. 18, no. 8, pp. 2121 - 2133, 2010.

[2] N. Hu, R.B. Dannenberg and G. Tzanetakis, "Polyphonic audio matching and alignment for music retrieval," in *Proc. WASPAA*, New Paltz, New York, USA, 2003, pp. 185-188.

This work was supported by NSF grant numbers IIS-0643752 and IIS-0812314.