Audio-visual Source Association for String Ensembles through Multi-modal Vibrato Analysis

Bochen Li, Chenliang Xu, Zhiyao Duan

University of Rochester

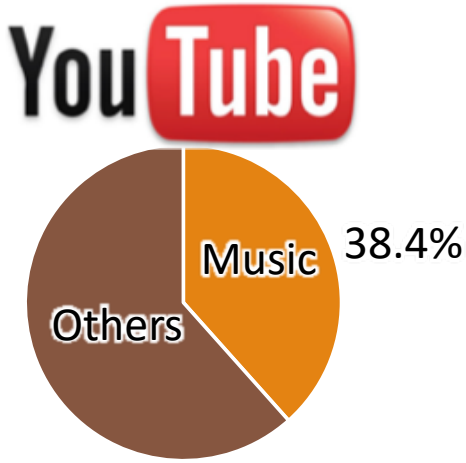
14th Sound and Music Computing Conference

July 5 – 8, 2017

Espoo, Finland

Background

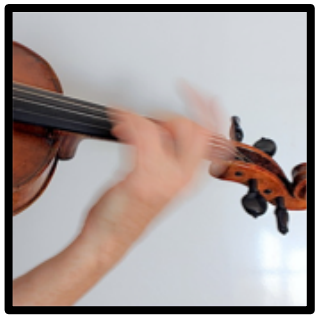
- Music → multi-modal art form
- See and listen → more enjoyment
- Popular music video streaming service



Background

Multi-modal MIR

- Instrument Recognition
- Playing Activity Detection
- Polyphonic Music Analysis
- Fingering Estimation
- Conductor Following

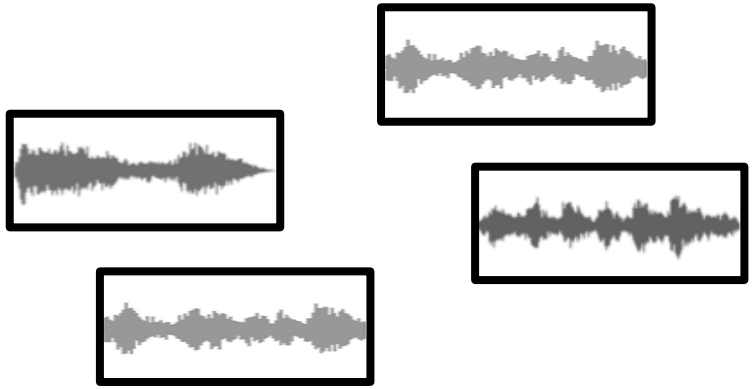


The Problem – Audio-visual Source Association

String Music Performance

Detected Players

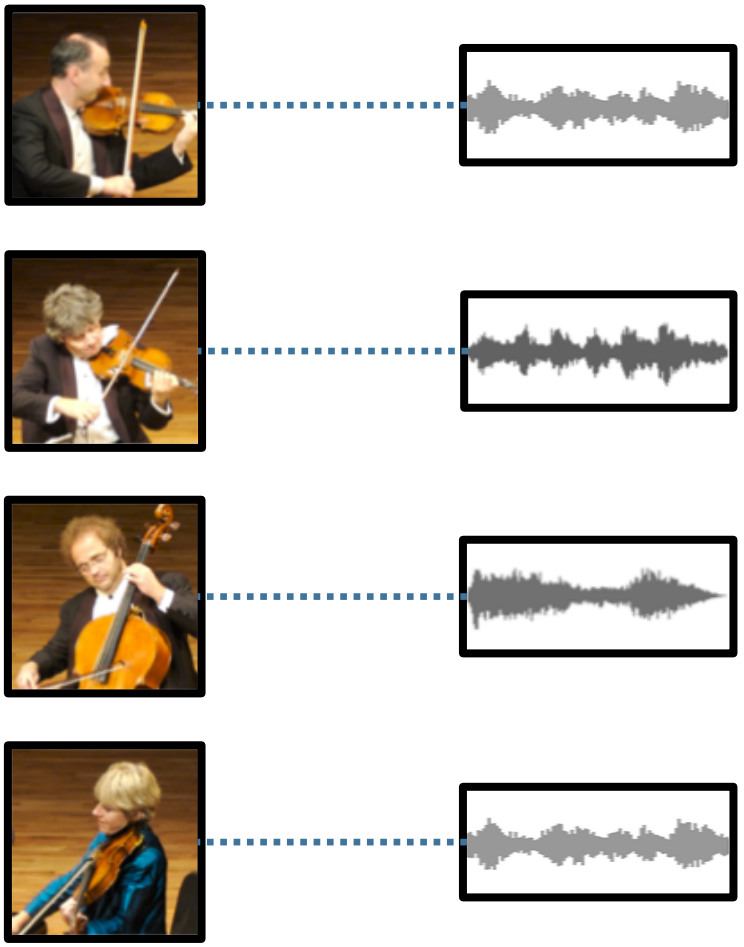
Separated Sound Tracks



The Problem – Audio-visual Source Association

String Music Performance

Audio-visual Source Association



The Problem – Audio-visual Source Association

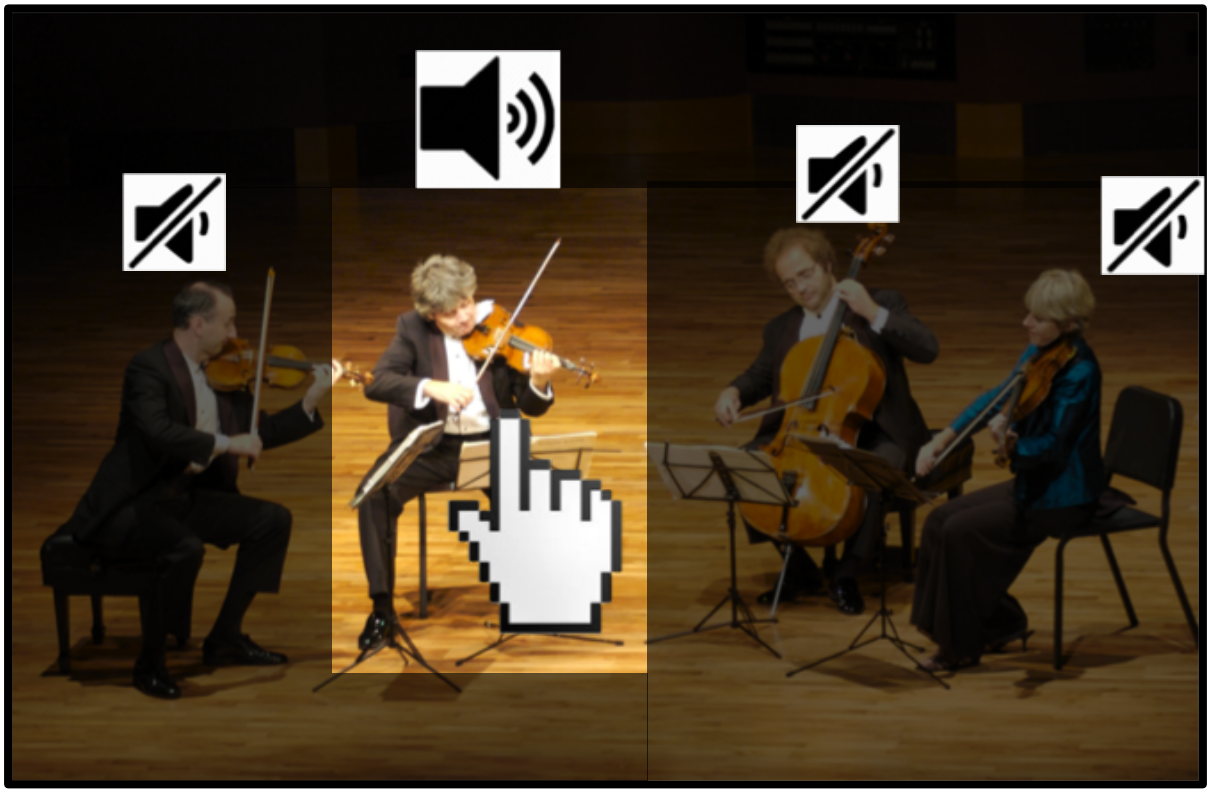
Application

- Intuitive and user-friendly interaction with music performance videos
- Smart Music Editor
- Concert cameras automatically take close-up shots of the leading player/instrument

The Problem – Audio-visual Source Association

Application

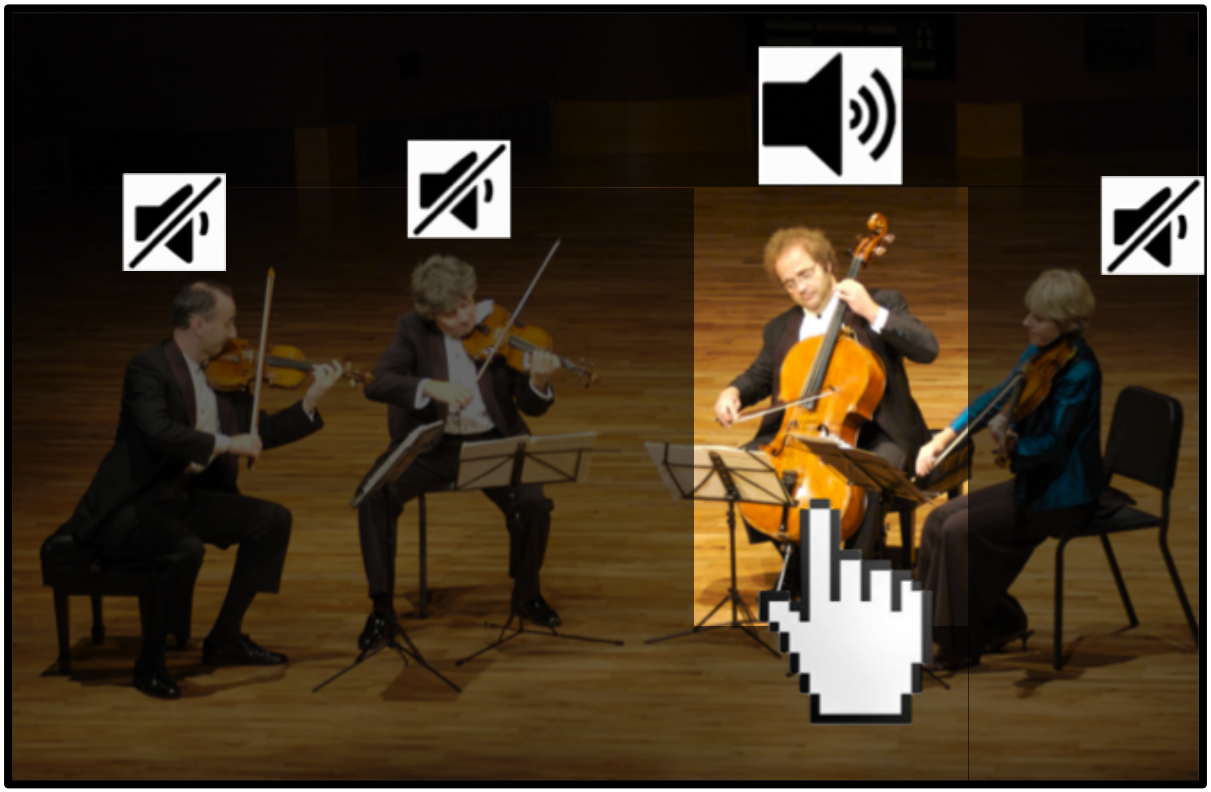
- Intuitive and user-friendly interaction with music performance videos
- Smart Music Editor
- Concert cameras automatically take close-up shots of the leading player/instrument



The Problem – Audio-visual Source Association

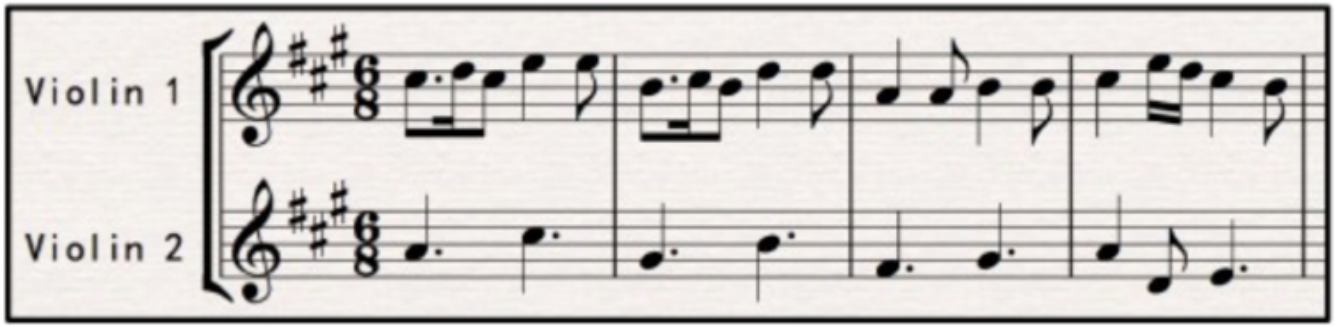
Application

- Intuitive and user-friendly interaction with music performance videos
- Smart Music Editor
- Concert cameras automatically take close-up shots of the leading player/instrument



Bow Motion Analysis

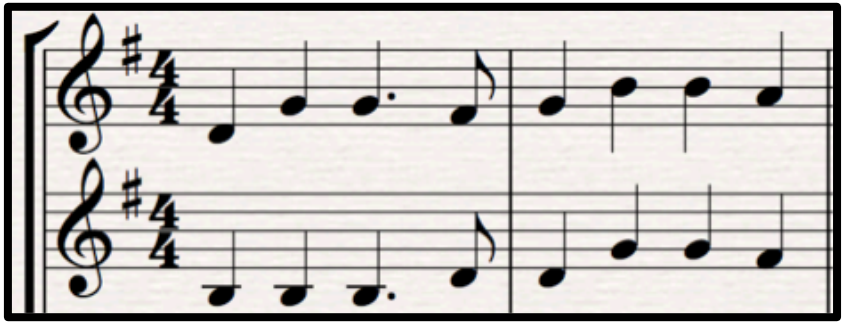
- Bow Motion <–> Note Onsets



Prior Work

Limitations

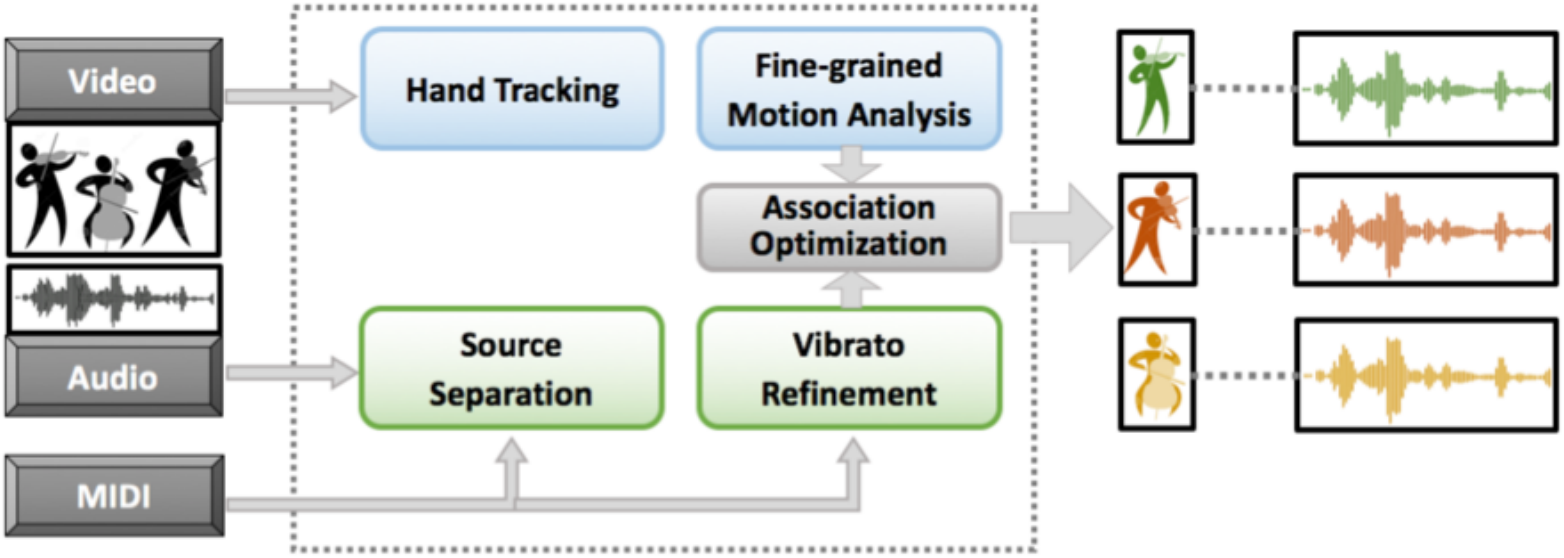
- When players have the same rhythm



Proposed System Overview

Vibrato Features for String Instruments

- Vibrato → Audio pitch fluctuations
- Vibrato → Fine motions of left hand
- Correlate pitch fluctuations with fine motions of left hand



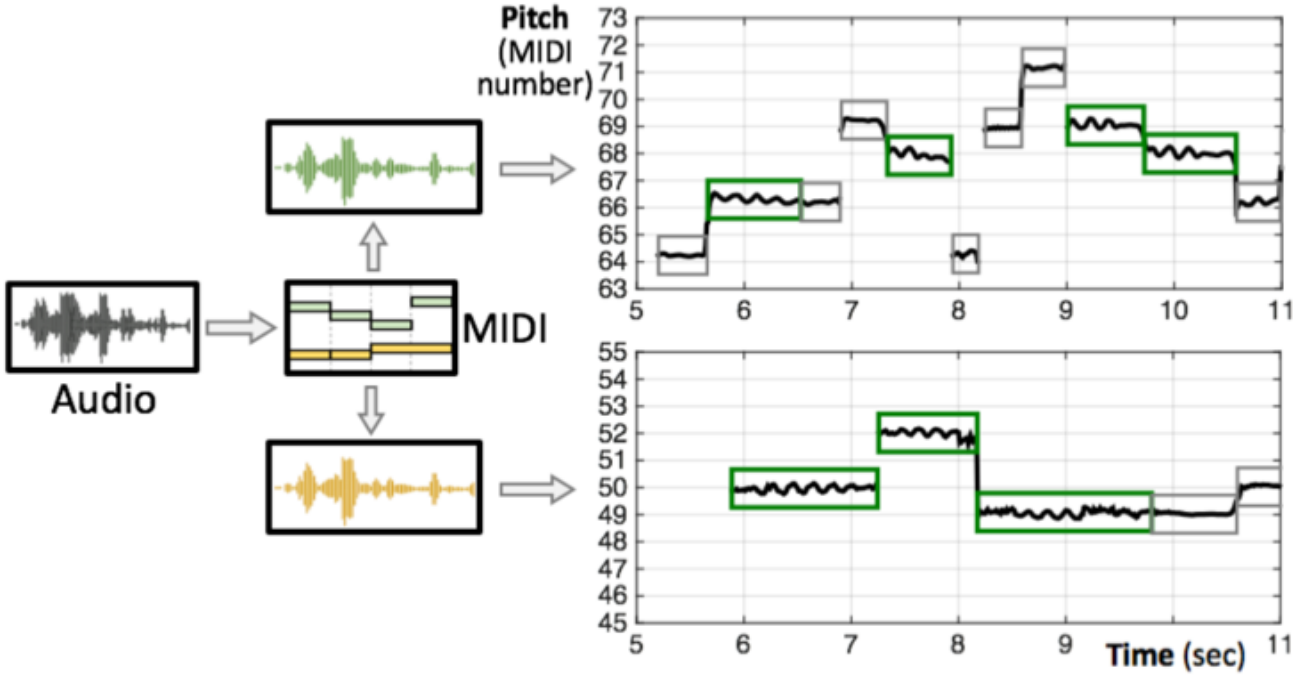
Method – Audio Analysis

Score-informed Source Separation

- Audio-score alignment
- Harmonic mask

Vibrato Extraction

- Score-informed pitch refinement on separated sources
- Auto-correlation on pitch trajectory

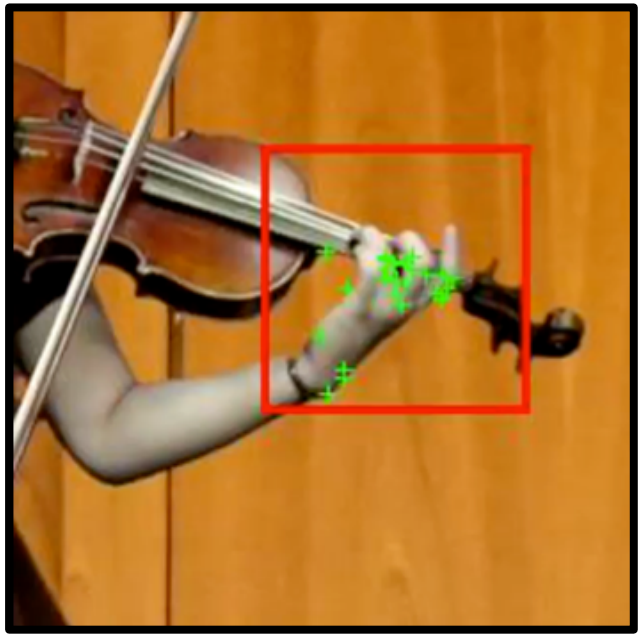


[2] Z. Duan and B. Pardo, "Soundprism: An online system for score-informed source separation of music audio," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp., 2011.

Method – Video Analysis

Hand Tracking

- Kanade-Lucas-Tomasi (KLT) tracker with 30 feature points
- Bounding box: 70*70 pixels, centered at the median position of feature points
- Re-initialize feature points every 20 frames



Method – Video Analysis

Fine-grained Motion Capture

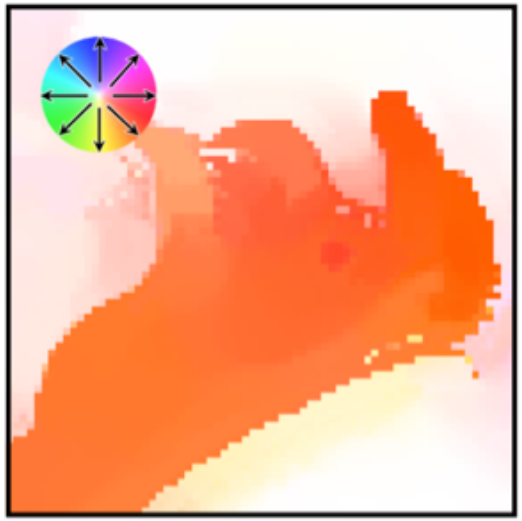
- Optical flow estimation → pixel-level motion velocities
- Average the motion velocities within the bounding box:

$$\mathbf{u}(t) = [u_x(t), u_y(t)]$$

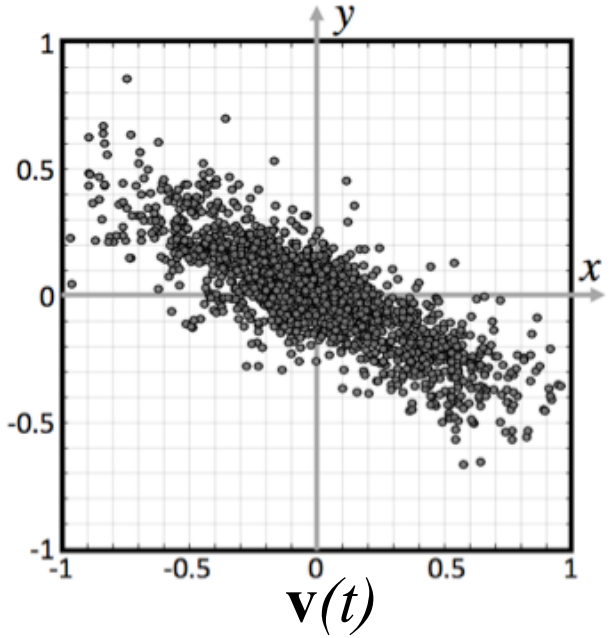
- Subtract its moving average to eliminate body motion:

$$\mathbf{v}(t) = \mathbf{u}(t) - \bar{\mathbf{u}}(t)$$

Original Frame



Color-encoded Optical Flow



$\mathbf{v}(t)$

Method – Video Analysis

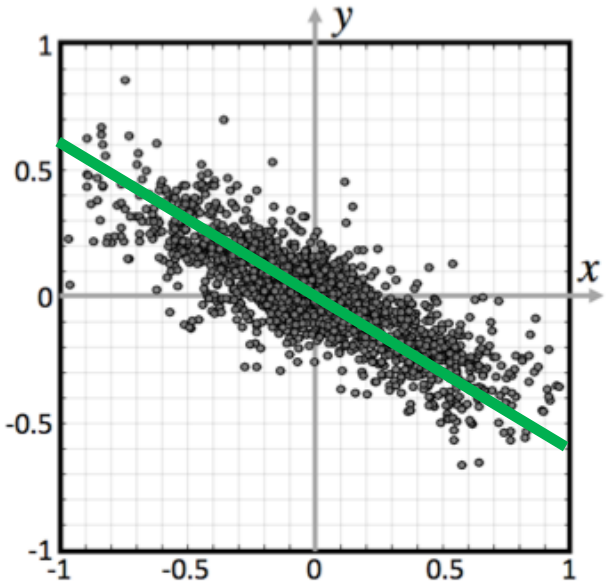
Fine-grained Motion Capture

- Principal Component Analysis (PCA)

→ Identify principal motion along the fingerboard

→ 1-D Motion Velocity Curve:

$$V(t) = \frac{\mathbf{v}(t)^T \tilde{\mathbf{v}}}{\|\tilde{\mathbf{v}}\|}$$

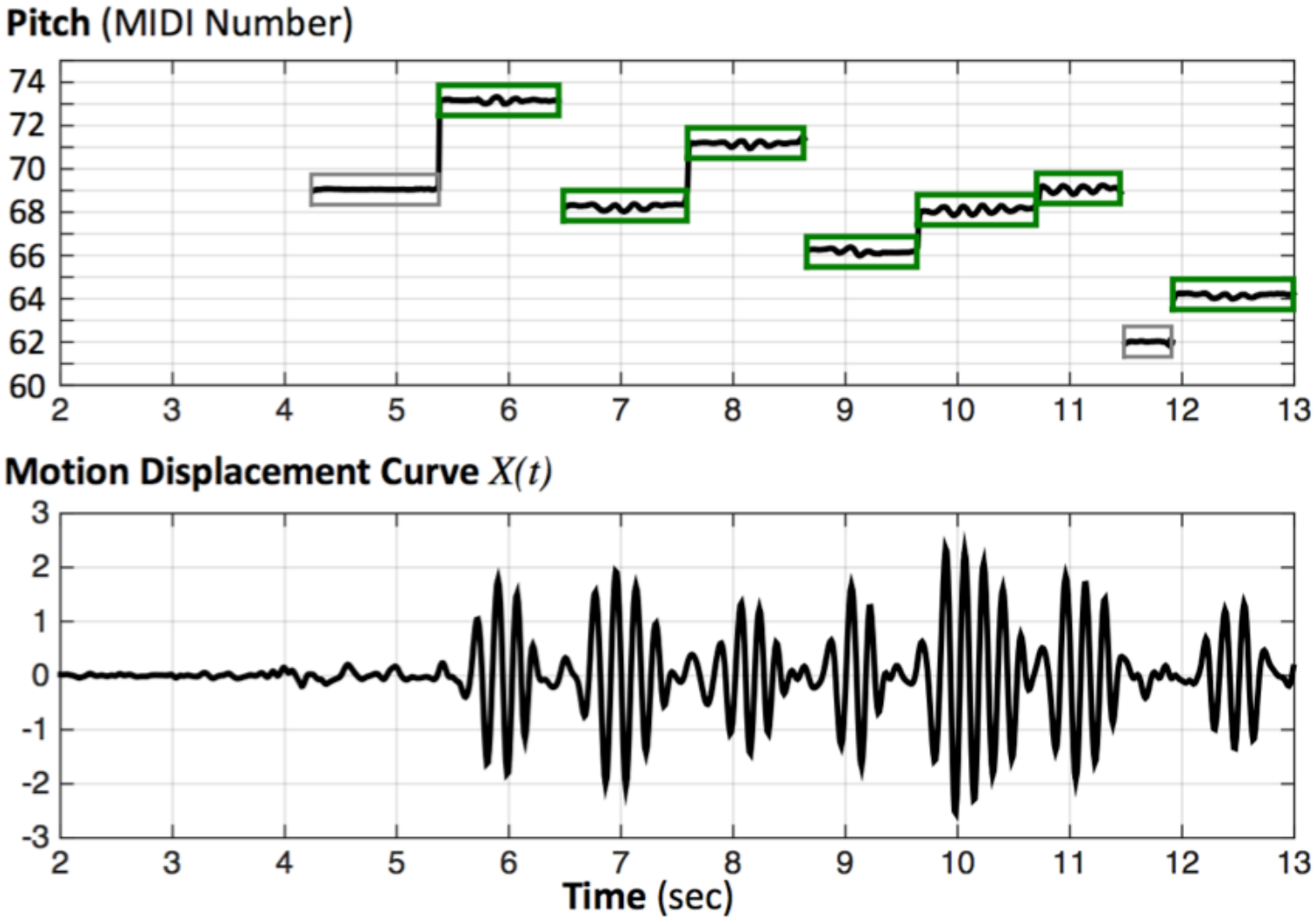


- Integration on $V(t)$ → Motion Displacement Curve:

$$X(t) = \int_0^t V(\tau) d\tau$$

Method – Video Analysis

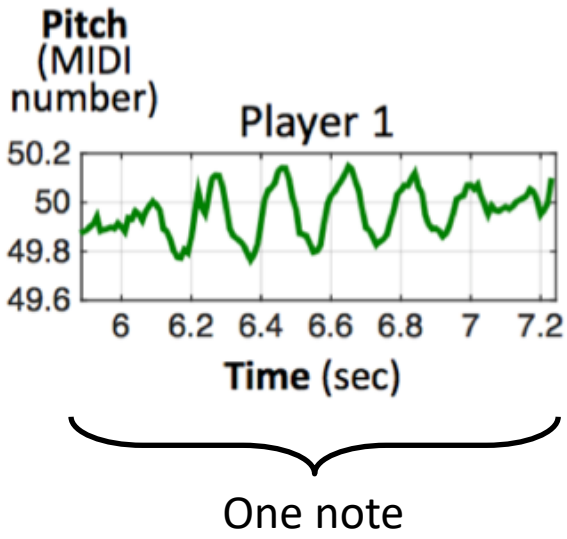
Fine-grained Motion Capture



Method – Source-player Association

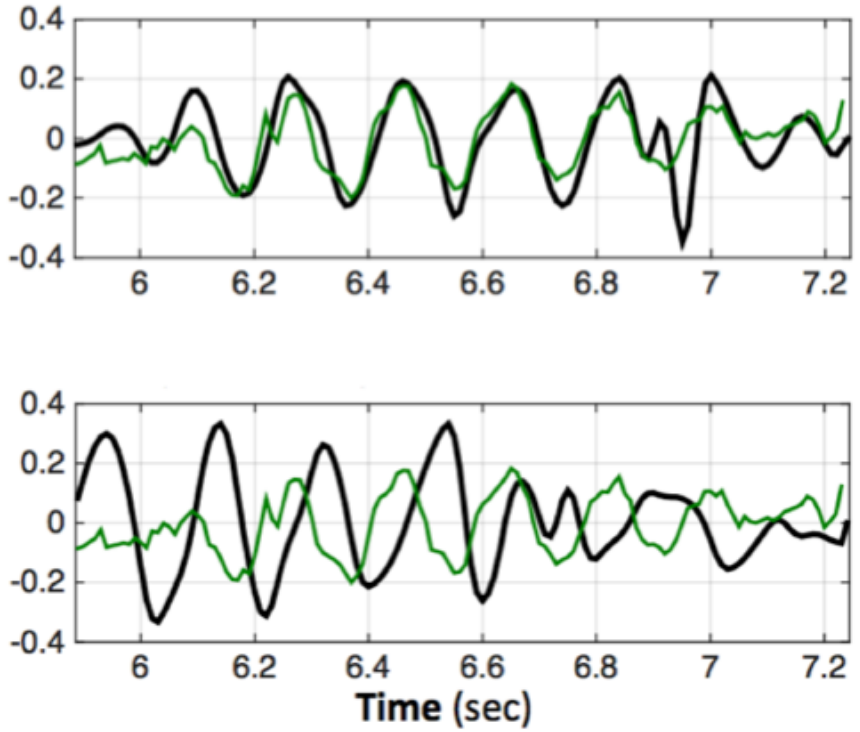
Motion Displacement Curve

Pitch Contour



Associated player

Pitch & Motion (normalized)



Not associated player

Method – Source-player Association

- Note-level matching score
  → Cross-correlation

Normalized pitch

Normalized motion

i-th note

$$m^{[p,q]}(i) = \exp\left\{ \frac{\left\langle \hat{F}^{[p]}(\mathbf{t}_i) \cdot \hat{X}^{[q]}(\mathbf{t}_i) \right\rangle}{\left\| \hat{F}^{[p]}(\mathbf{t}_i) \right\| \left\| \hat{X}^{[q]}(\mathbf{t}_i) \right\|} \right\}$$

Audio track index

Player index

- Track-level matching score
  → Sum of note-level matching score

Total number of vibrato notes in the p-th track

$$M^{[p,q]} = \sum_{i=1}^{N_{\text{vib}}^{[p]}} m^{[p,q]}(i)$$

Method – Source-player Association

- Association score

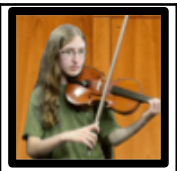
Total number of tracks (i.e., players)

Track-level matching score

$$S_\sigma = \prod_{p=1}^{K} M^{[p,\sigma(p)]}$$

One permutation

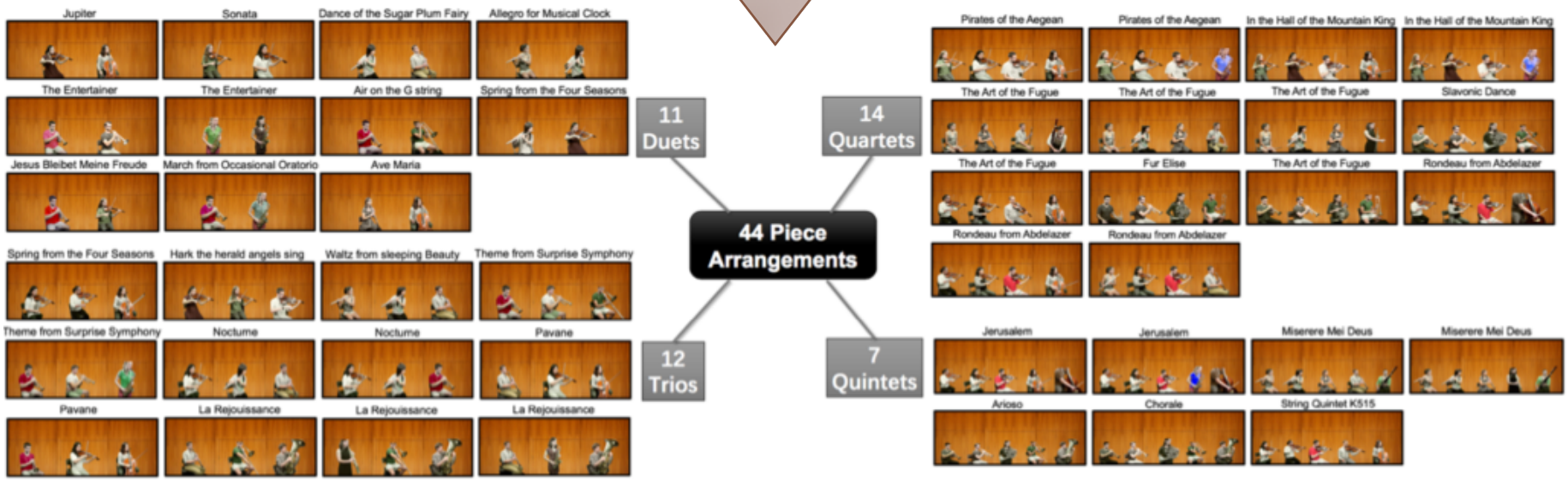
Output the permutation that maximizes the association score

				
♫	$M_{1,1}$	$M_{2,1}$	$M_{3,1}$	$M_{4,1}$
♫	$M_{1,2}$	$M_{2,2}$	$M_{3,2}$	$M_{4,2}$
♫	$M_{1,3}$	$M_{2,3}$	$M_{3,3}$	$M_{4,3}$
♫	$M_{1,4}$	$M_{2,4}$	$M_{3,4}$	$M_{4,4}$

Experiments

Dataset: URMP Dataset [3]

- Individually recorded and assembled together
- 14 instruments, 44 piece arrangements



[3] B. Li *, X. Liu *, K. Dinesh, Z. Duan, and G. Sharma, "Creating a musical performance dataset for multimodal music analysis: Challenges, insights, and applications," *IEEE Trans. Multimedia*, under review.

Experiments

Piece Selection

- 19 pieces → 5 duets, 4 trios, 7 quartets, 3 quintets
- Selection criteria: contains at most 1 non-string instrument
- Same set as the baseline system (bow motion ←→ note onset)

Evaluation Measure

- **Note-level Matching Accuracy:**

  The % of vibrato notes that are best matched to the correct player, according to the note-level matching score

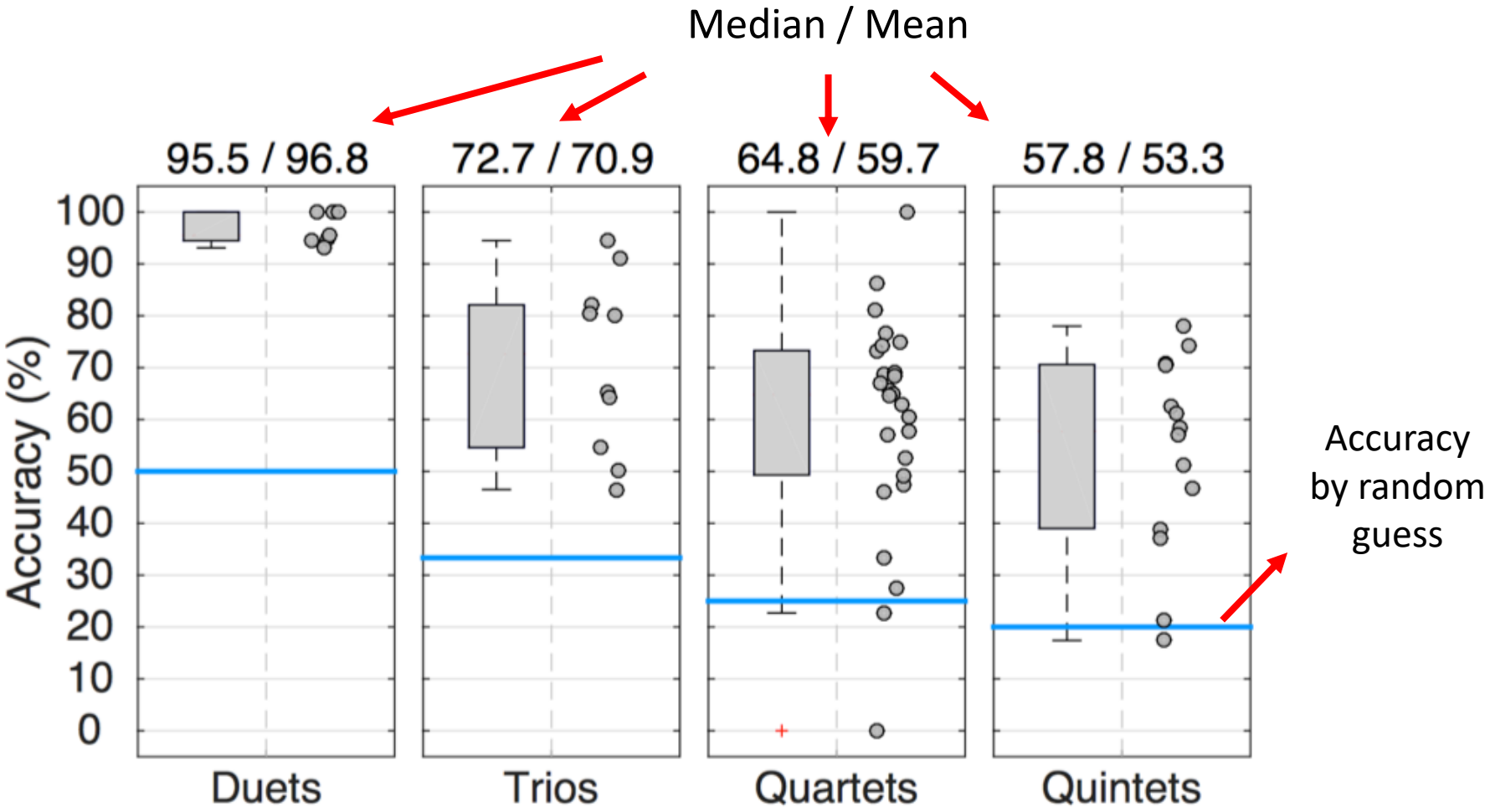
- **Piece-level Association Accuracy:**

  The % of pieces that the correct association is returned, according to the piece-level association score

(Polyphony increases → Number of error candidates increases in factorial rate)

Experiments

Results: Note-level Matching Accuracy



Experiments

Results: Piece-level Association Accuracy

Metadata				Association Measures	
No.	Dataset Folder Name (with Instrument Types)	Piece Length (mm:ss)	Polyphony - (No. Permutations)	No. Correctly Associated Sources	Rank of Correct Association
1	01_Jupiter_vn_vc	01:03	2 - (2)	2	1
2	02_Sonata_vn_vn	00:46	2 - (2)	2	1
3	08_Spring_fl_vn	00:35	2 - (2)	2	1
4	09_Jesus_tpt_vn	03:19	2 - (2)	2	1
5	11_Maria_ob_vc	01:44	2 - (2)	2	1
6	12_Spring_vn_vn_vc	02:11	3 - (6)	3	1
7	13_Hark_vn_vn_va	00:47	3 - (6)	3	1
8	19_Pavane_cl_vn_vc	02:13	3 - (6)	1	2
9	20_Pavane_tpt_vn_vc	02:13	3 - (6)	3	1
10	24_Pirates_vn_vn_va_vc	00:50	4 - (24)	4	1
11	25_Pirates_vn_vn_va_sax	00:50	4 - (24)	4	1
12	26_King_vn_vn_va_vc	01:25	4 - (24)	4	1
13	27_King_vn_vn_va_sax	01:25	4 - (24)	2	1
14	32_Fugue_vn_vn_va_vc	02:54	4 - (24)	4	1
15	35_Rondeau_vn_vn_va_db	02:08	4 - (24)	4	1
16	36_Rondeau_vn_vn_va_vc	02:08	4 - (24)	4	1
17	38_Jerusalem_vn_vn_va_vc_db	01:59	5 - (120)	5	1
18	39_Jerusalem_vn_vn_va_sax_db	01:59	5 - (120)	5	1
19	44_K515_vn_vn_va_va_vc	03:45	5 - (120)	5	1

- Overall Accuracy: 94.7% (18 out of 19)
  Compared with Baseline: 89.5% (based on bow motion/audio onset)
- Error Case: No vibrato is used in the performance

Conclusions & Future Work

Conclusions

- **Audio-visual source association for string music, by correlating pitch fluctuations and left-hand motions**
- **Highly effective, not demanding on camera angles**
- **Limitations: Vibrato is not guaranteed to appear in all pieces**

Future Work

- **Combine all motion features in string music**

  Bow & Vibrato & Body movement & …

- **Video → Vibrato analysis (rate & extent)**

  From monophonic to polyphonic

- **Step into woodwind & brass instruments**
- **Audio-visual Source Separation**