Visually Informed Multi-Pitch Analysis of String Ensembles

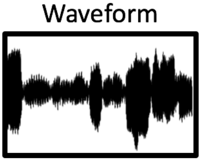
Karthik Dinesh, Bochen Li, Xinzhao Liu, Zhiyao Duan, Gaurav Sharma

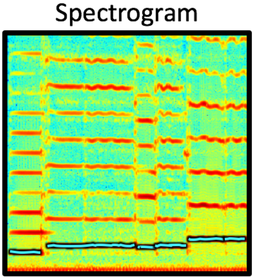
Department of Electrical and Computer Engineering, University of Rochester

March 9, 2017

Pitch in Music

- Pitch - fundamental frequency of musical note from an instrument



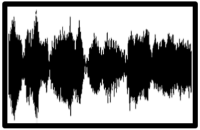


- Pitch changes with time as notes and vibrato change

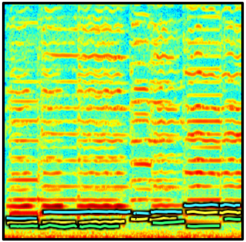
Multi-pitch Analysis

- Multiple music instrument ensemble has pitches corresponding to notes from each instrument - multiple pitches

Waveform



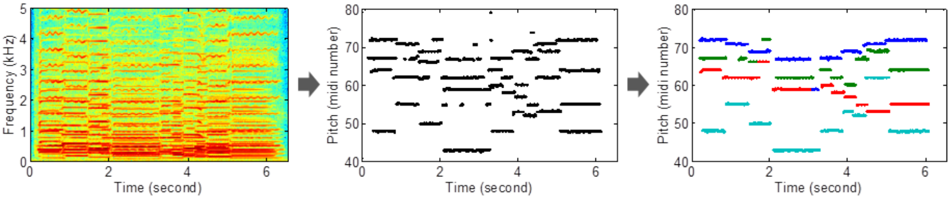
Spectrogram



Introduction: Multi-pitch Estimation and Streaming

Multi-pitch Analysis

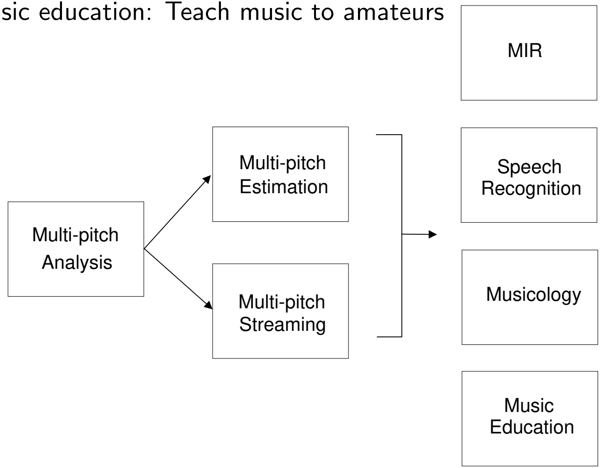
- Multi-pitch Estimation (MPE): Estimate instantaneous pitches and polyphony



- Multi-pitch Streaming (MPS): Organize the estimated pitches into streams corresponding to individual sound sources

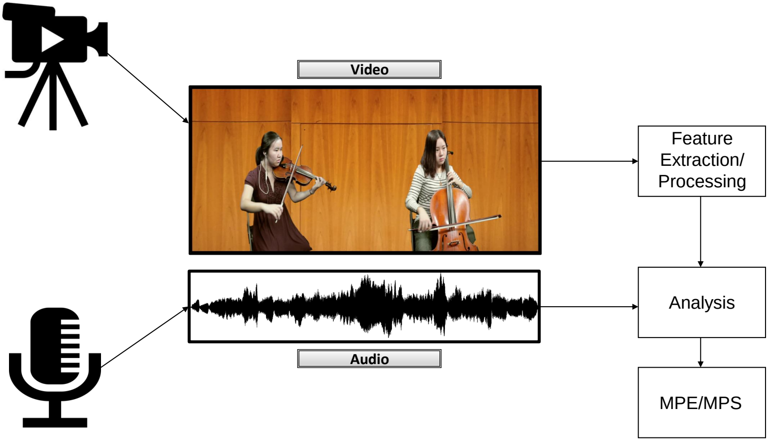
Introduction: Applications of Multi-pitch Analysis

- Multipitch analysis
  - MIR: Music transcription, source separation, melody extraction
  - Speech recognition: Multi-talk recognition, prosody analysis
  - Musicology: Scholarly analysis
  - Music education: Teach music to amateurs



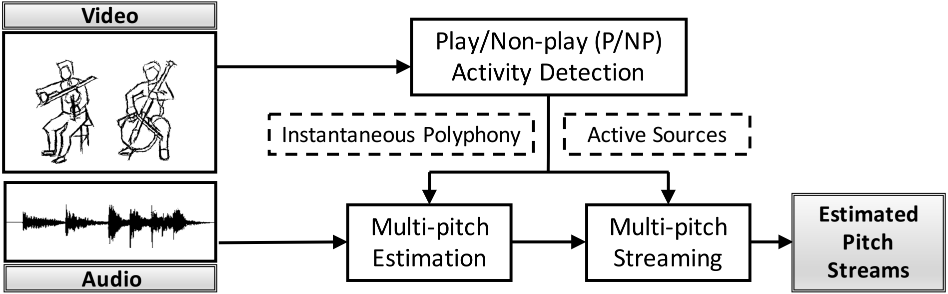
Contribution: Augmenting MPE/MPS with Video

- MPE/MPS based on audio alone challenging
- Video modality provides valuable information
- Multimedia research has gained prominence
- Limited video informed work till date

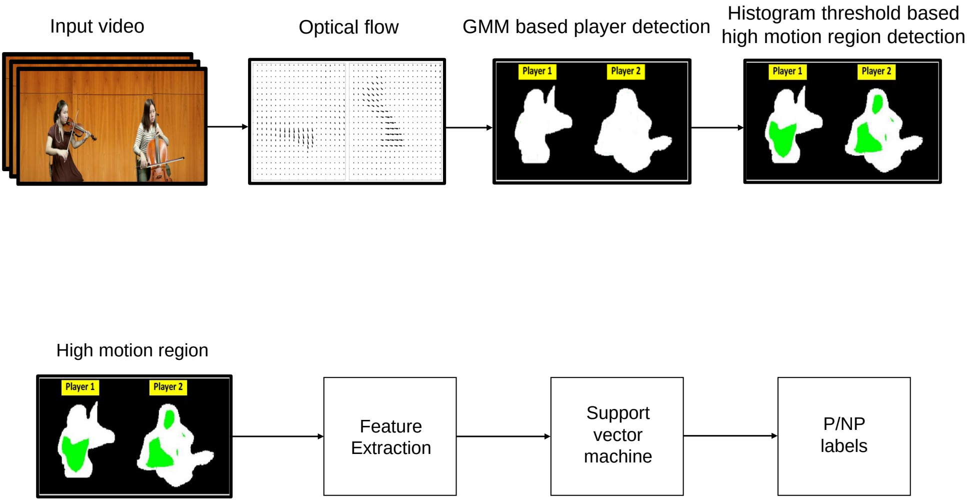


Visually Informed Multi-pitch Analysis: Framework

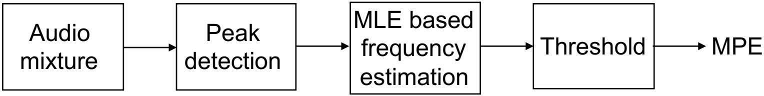
- Video module $\rightarrow$ play/non-play (P/NP) activity
- P/NP activity $\rightarrow$ instantaneous polyphony (for MPE)
  helps organize pitches to active sources (for MPS)



P/NP detection: Framework



Multipitch Analysis: Prior Audio-Only Multi-pitch Estimation [2]



$$\hat{\theta} = \underset{\theta \in \boldsymbol{\Theta}}{arg\ max} \pounds(\mathbf{O} \mid \theta)$$

$$\pounds(\theta) = \pounds_{peak}(\theta) . \pounds_{non-peak}(\theta)$$

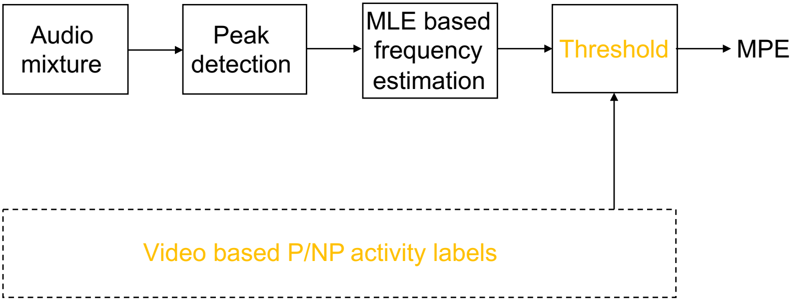
$\theta \rightarrow Set\ of\ fundamental\ frequency$

$\mathbf{O} \rightarrow Obs.\ from\ power\ spectrum$

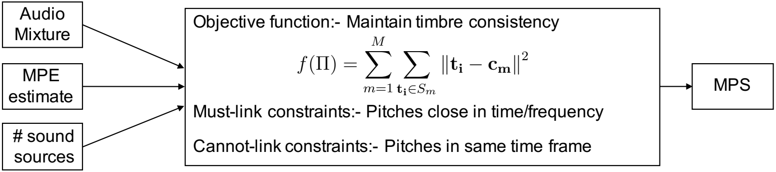
$\boldsymbol{\Theta} \rightarrow space\ of\ possible\ sets\ of\ \theta$

Multipitch Analysis: Video Based Multi-pitch Estimation

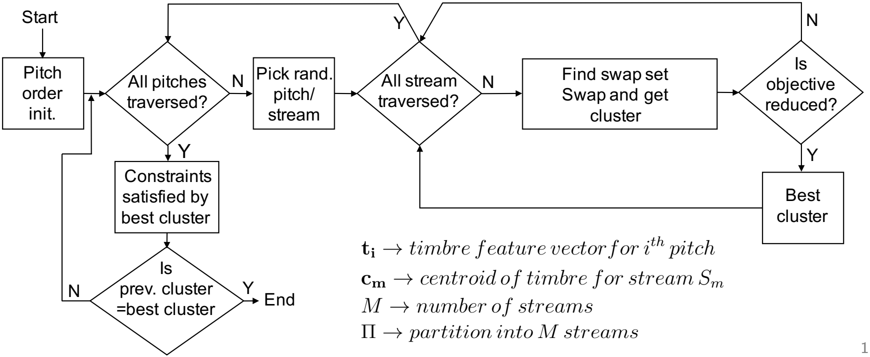
- P/NP labels inform instantaneous polyphony
- Instantaneous polyphony used as threshold



Multipitch Analysis: Prior Audio-Only Multi-pitch Streaming [3]



Constrained Clustering Algorithm



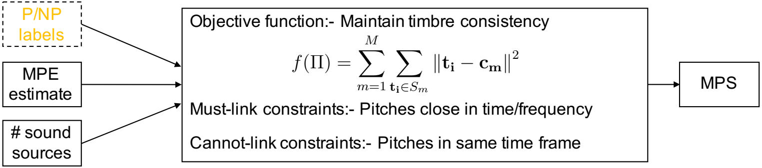
$\mathbf{t_i} \rightarrow timbre\ feature\ vector\ for\ i^{th}\ pitch$

$\mathbf{c_m} \rightarrow centroid\ of\ timbre\ for\ stream\ S_m$

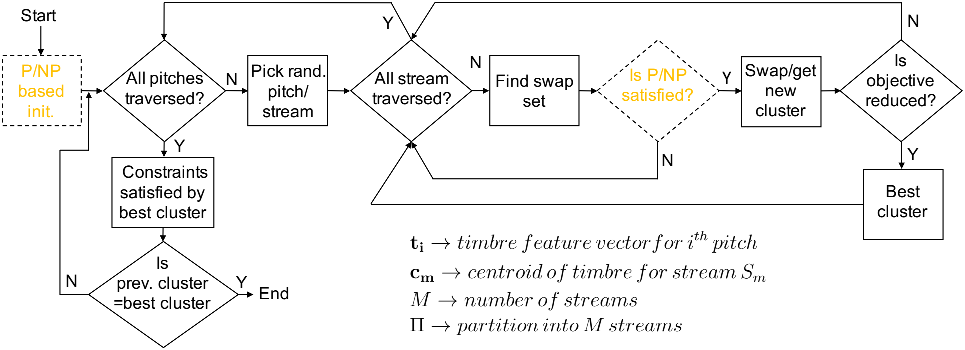
$M \rightarrow number\ of\ streams$

$\Pi \rightarrow partition\ into\ M\ streams$

Multipitch Analysis: Video Based Multi-pitch Streaming



Constrained Clustering Algorithm



$\mathbf{t_i} \rightarrow timbre\ feature\ vector\ for\ i^{th}\ pitch$

$\mathbf{c_m} \rightarrow centroid\ of\ timbre\ for\ stream\ S_m$

$M \rightarrow number\ of\ streams$

$\Pi \rightarrow partition\ into\ M\ streams$

Experimental Results

- Assessment on subset of URMP ground-truth dataset [4]
  - Focus on string ensembles including violin, viola, cello, and bass
  - 11 videos featuring 3 duets, 2 trios, 4 quartets, and 2 quintets
- P/NP SVM classifier used with radial basis function (RBF) kernel
- P/NP evaluation: leave one out cross validation error




Experimental Results: Performance Metrics

- P/NP detection accuracy:

$$P/NP\ detection\ acc = \frac{\#corr\ predicted\ labels\ w.r.t\ ground\ truth}{\#labels}$$

- MPE accuracy:

$$MPE\ acc = \frac{\#corr\ est\ pitch}{\#est\ pitch + \#gt\ pitch - \#corr\ est\ pitch}$$

- MPS accuracy:

$$MPS\ acc = \frac{\#corr\ est\ \&\ str\ pitch}{\#corr\ est\ \&\ str\ pitch + \#pitch\ in\ est\ not\ gt + \#pitch\ in\ gt\ not\ est}$$

corr $\rightarrow$ correct, est$\rightarrow$estimated, str$\rightarrow$streamed,gt$\rightarrow$ground truth

Experimental Results: P/NP Detection and MPE Accuracy

Piece No.	P/NP Detection Accuracy (%)					MPE Accuracy (%)		
	P1	P2	P3	P4	P5	Audio	Video PNP	GT PNP
# 1	97.4	91.5	-	-	-	70.2	83.6	85.1
# 2	93.6	93.3	-	-	-	68.7	72.2	74.2
# 3	81.1	71.3	-	-	-	58.5	62.7	70.0
# 4	92.5	91.4	78.4	-	-	59.8	65.9	68.6
# 5	93.9	92.9	89.4	-	-	75.0	76.7	79.0
# 6	83.4	88.4	78.6	73.4	-	49.5	52.3	56.3
# 7	69.3	73.6	75.1	70.1	-	52.1	52.0	59.0
# 8	90.0	90.9	84.6	86.4	-	62.2	62.3	66.6
# 9	93.1	95.5	82.4	91.5	-	62.2	63.3	65.7
# 10	91.9	92.3	88.5	94.1	91.2	47.4	52.3	53.3
# 11	74.2	75.1	70.0	75.3	62.5	46.4	44.0	48.8

Table: Results of video-based Play/Non-play detection and MPE accuracy of the 11 test pieces.

Experimental Results: Comparison of MPE Accuracies Audio/Video/Ground Truth

- Experiments on 53 duets, 38 trios and 14 quartets

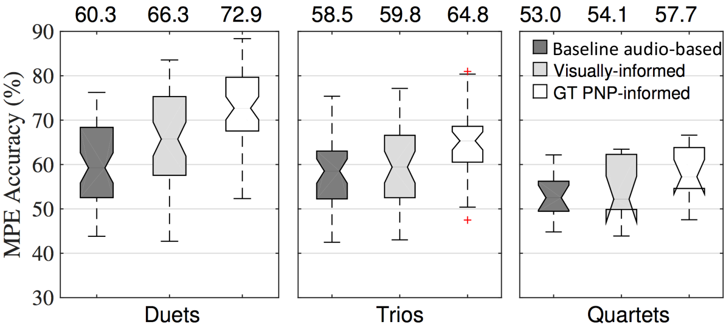


Figure: Boxplot of MPE accuracy grouped by polyphony on all subsets derived from the 11 pieces, comparing the baseline audio-based method (dark gray), proposed visually informed method (light gray), and the incorporation of ground-truth PNP labels (white). The number above each box shows the mean value of the box.

Experimental Results: Comparison of MPS Accuracies Audio/Video/Ground Truth

- Experiments on 53 duets, 38 trios and 14 quartets

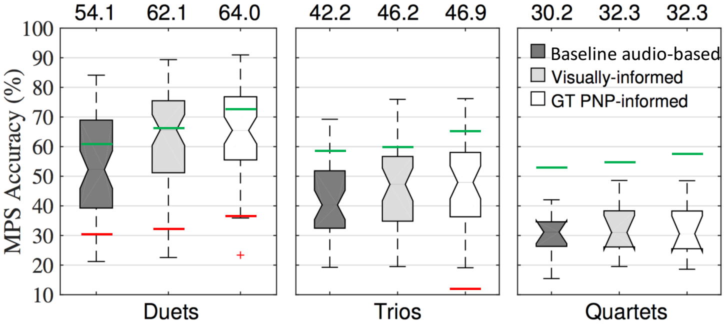


Figure: Boxplot of MPS accuracy grouped by polyphony on all subsets derived from the 11 pieces, comparing the baseline audio-based method (dark gray), proposed visually informed method (light gray), and the incorporation of ground-truth PNP labels (white). The number above each box shows the mean value of the box.

Conclusion

- We demonstrated a novel technique of visually informed multi-pitch analysis for string ensembles
- Video based play/non-play detection technique was used
  - To obtain concurrent pitches in each time frame (MPE)
  - To assign the estimated pitches to corresponding sound sources (MPS)
- Experimental results show
  - Video based P/NP detection has accuracy of 85.3%
  - Statistically significant improvement on both the MPE and MPS accuracy at a significance level of 0.01 in most cases
- With improvement in underlying MPE/MPS integration with P/NP, better results can be obtained

References

[1] D. Sun, S. Roth, and M. J. Black, “Secrets of optical flow estimation and their principles,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2432–2439.

[2] Z. Duan, B. Pardo, and C. Zhang, “Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 8, pp. 2121–2133, 2010.

[3] Z. Duan, J. Han, and B. Pardo, “Multi-pitch streaming of harmonic sound mixtures,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 1, pp. 138–150, 2014.

[4] 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*, submitted. Available: https://arxiv.org/abs/1612.08727.

Back Up Slides

Pieces Used in Experiments

Piece number	Piece Name	Polyphony	Performance Style Description
#1	01_Jupiter_vn_vc	2	Motion is easy to capture. All players are playing at most time
#2	02_Sonata_vn_vn	2	Motion is easy to capture. All players are playing at most time
#3	19_Pavane_cl_vn_vc	3	Some plucking motion for the violin and cello
#4	12_Spring_vn_vn_vc	3	Motion is easy to capture for player 1 and 2. For player 3, some soft articulation is from slow motion, which may be difficult to capture
#5	13_Hark_vn_vn_va	3	Motion is easy to capture. All players are playing at most time
#6	24_Pirates_vn_vn_va_vc	4	Motion is easy to capture. All players are playing at most time
#7	26_King_vn_vn_va_vc	4	A lot of repeated notes, where the bow motion is slight
#8	32_Fugue_vn_vn_va_vc	4	Motion is easy to capture. Different players play alternatively sometimes
#9	36_Rondeau_vn_vn_va_vc	4	Motion is easy to capture. All players are playing at most time.
#10	38_Jerusalem_vn_vn_va_vc_db	5	Motion is easy to capture. All players are playing at most time.
#11	44_K515_vn_vn_va_va_vc	5	Some fast notes are played by legato bowing, where the bow motion is slow.

Table: Pieces used in the experiment with polyphony and performance style

Problematic Pieces

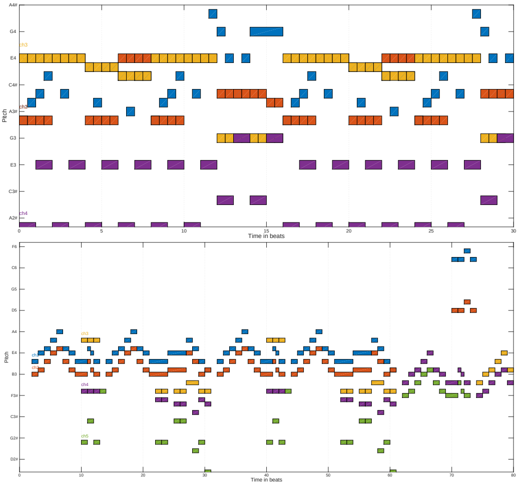
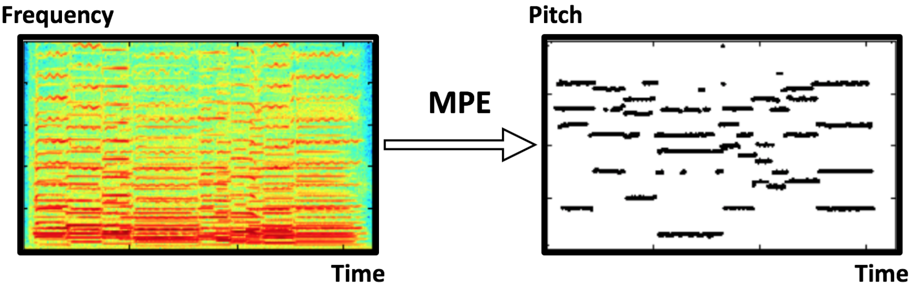


Figure: MIDI plot for segments from pieces (#7) 26-In hall of mountain king (top) and (#11) 44-K515 (bottom) which have limited bow motion

Multi-pitch Estimation

- Likelihood method [2]
- Model peak/non-peak region of spectrum
- Interative greedy search → estimate pitch one by one



Audio Based Multipitch Analysis

Multi-pitch Streaming

- Constrained clustering method [3]
- Constraints on timbre consistency
- Constraints on time-frequency relationship