Harmonically Informed Multi-pitch Tracking

Zhiyao Duan, Jinyu Han and Bryan Pardo
EECS Dept., Northwestern Univ.
Interactive Audio Lab, http://music.cs.northwestern.edu

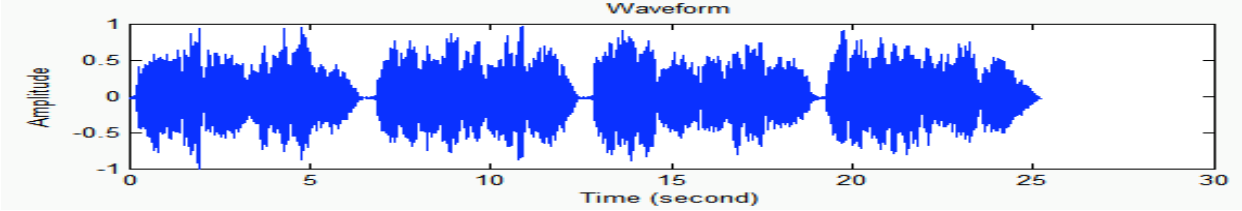
For presentation in ISMIR 2009, Kobe, Japan.

NORTHWESTERN
UNIVERSITY

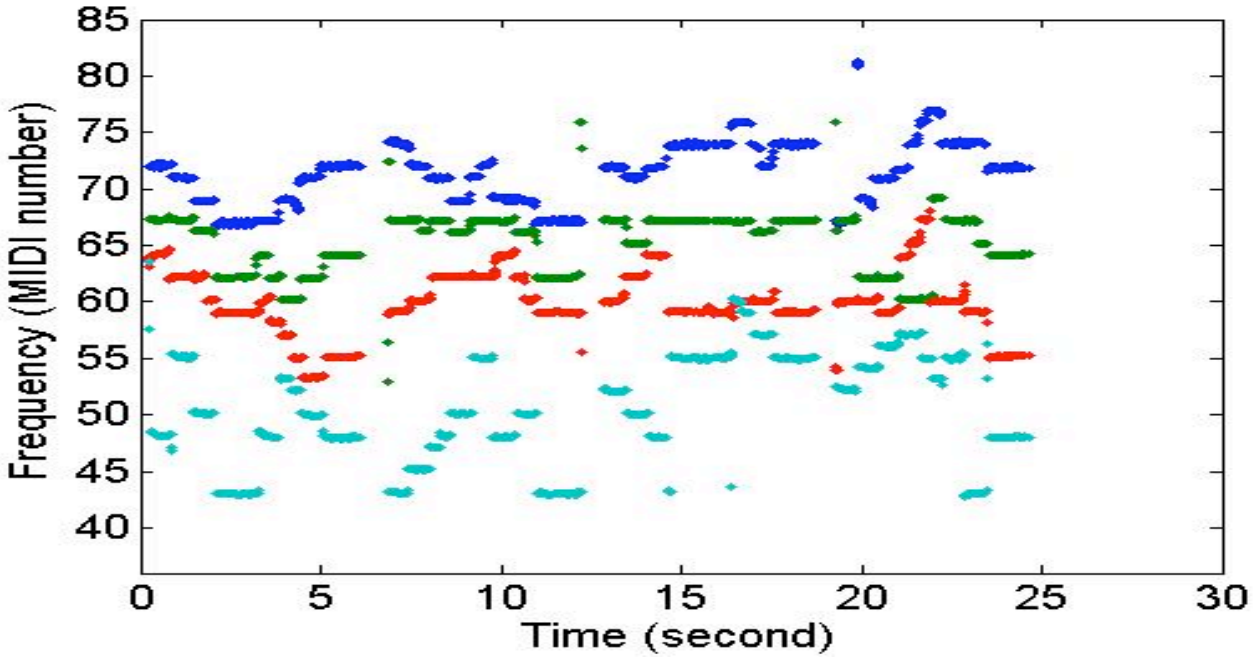
interactive
audio lab

The Multi-pitch Tracking Task

- Given polyphonic music played by several monophonic harmonic instruments



- Estimate a pitch trajectory for each instrument

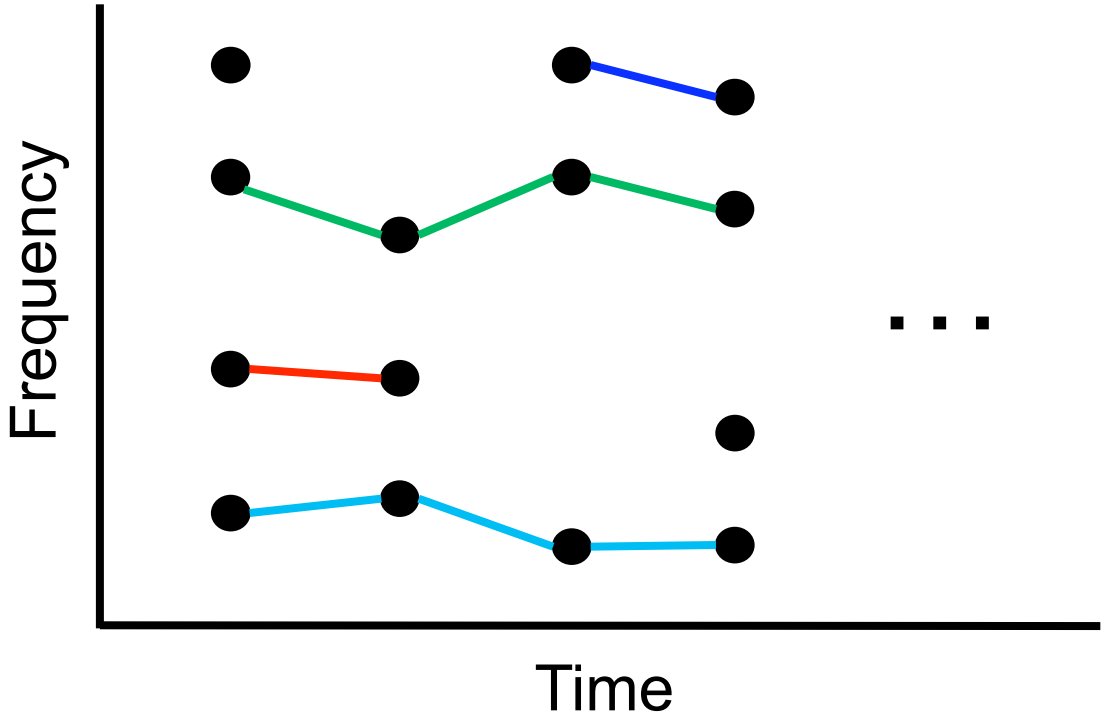


Potential Applications

- Automatic music transcription
- Harmonic source separation
- Other applications
  - Melody-based music search
  - Chord recognition
  - Music education
  - ……

The 2-stage Standard Approach

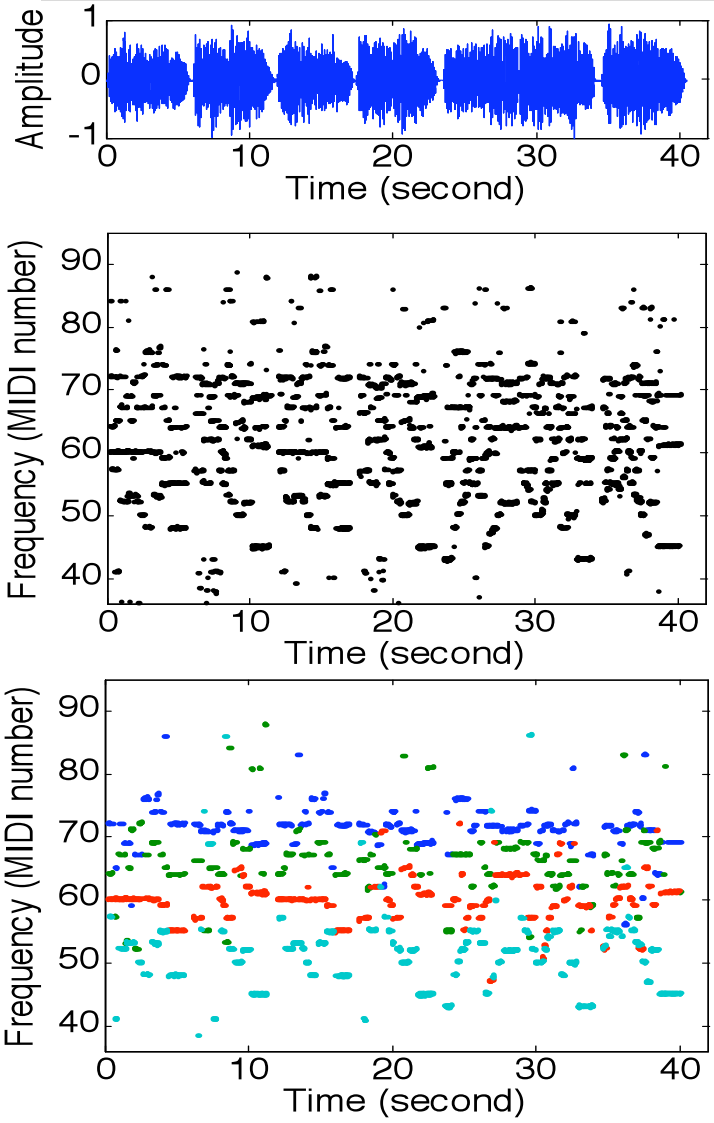
- Stage 1: Multi-pitch Estimation (MPE) in each single frame
- Stage 2: Connect pitch estimates across frames into pitch trajectories

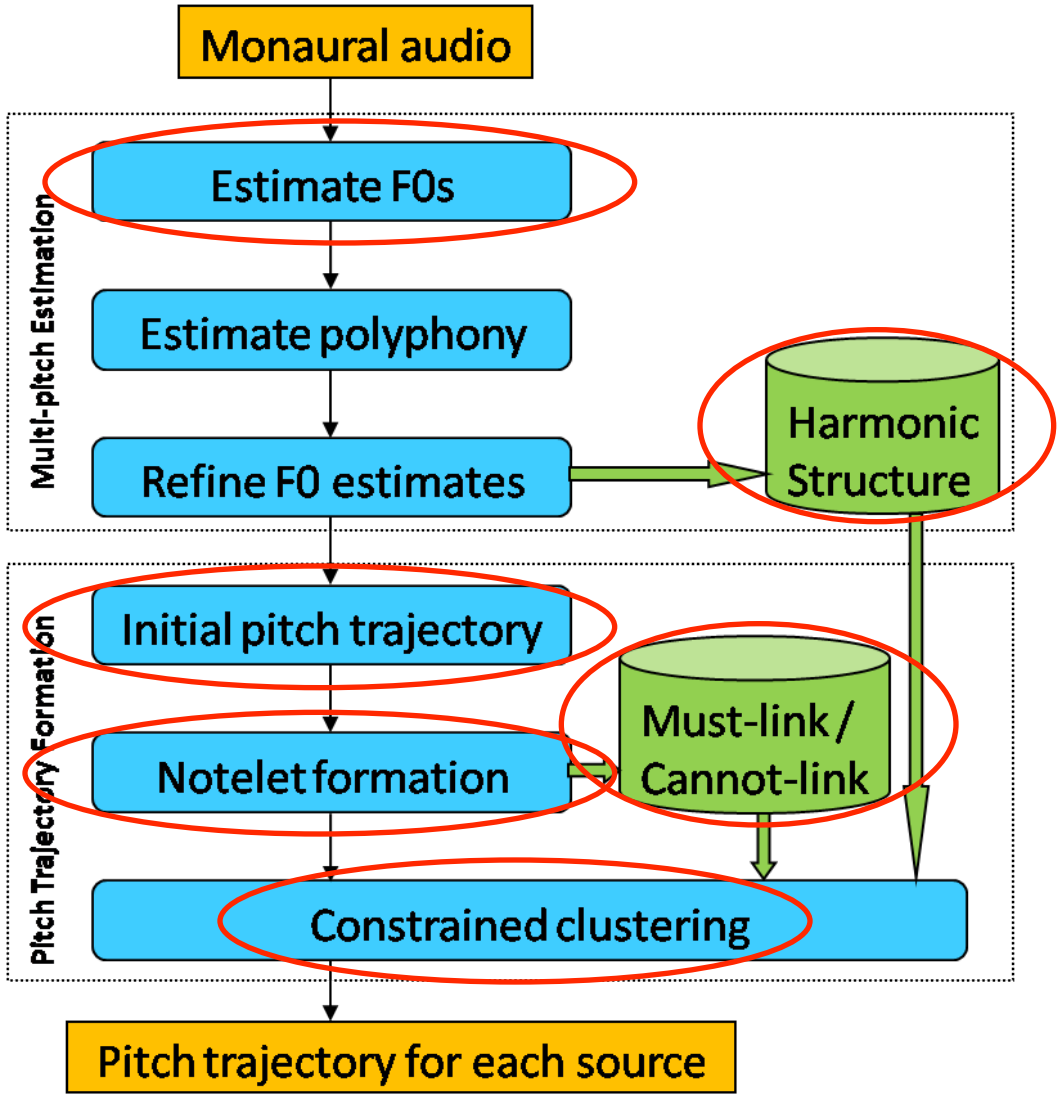


State of the Art

- How far has existing work gone?
  - MPE is not very robust
  - Form short pitch trajectories (within a note) according to local time-frequency proximity of pitch estimates
- Our contribution
  - A new MPE algorithm
  - A constrained clustering approach to estimate pitch trajectories across multiple notes

System Overview





Multi-pitch Estimation in Single Frame

- A maximum likelihood estimation method

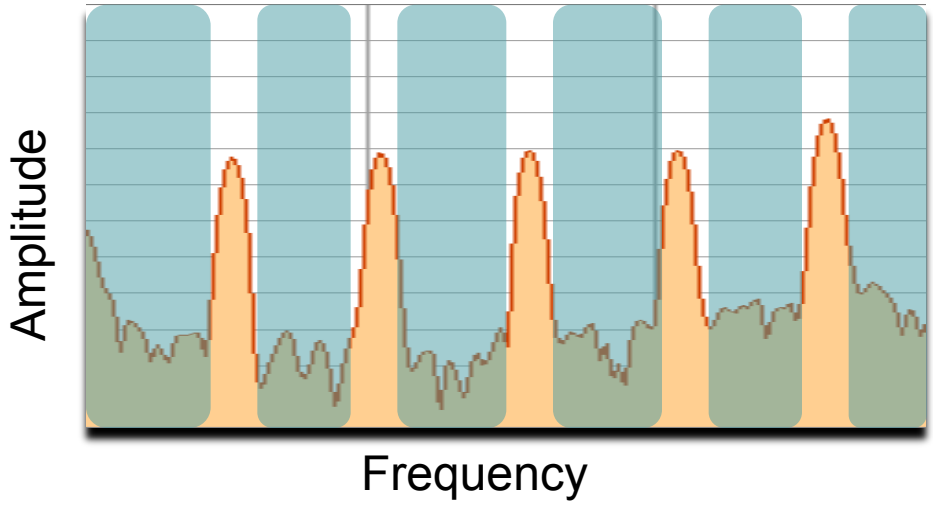
$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{L}\left(\mathbf{O}|\boldsymbol{\theta}\right)$$

Best F0 estimate
(a set of F0s)

Observed power
spectrum

F0 hypothesis,
(a set of F0s)

- Spectrum: peaks & the non-peak region

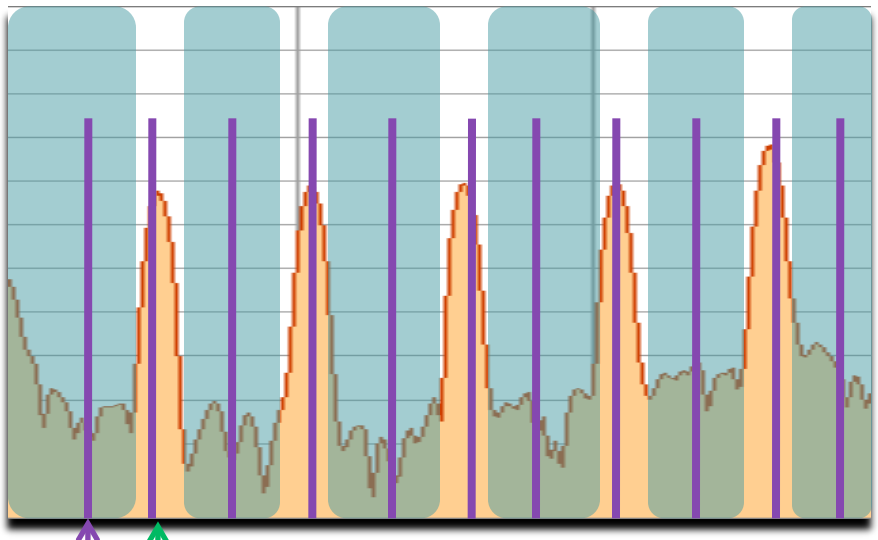


Likelihood Definition

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{peak}}(\theta) \cdot \mathcal{L}_{\text{non-peak region}}(\theta)$$

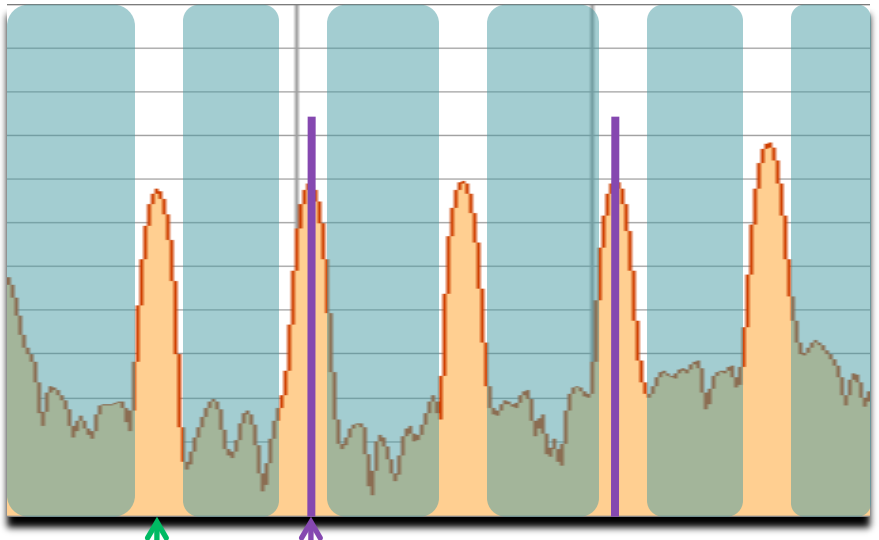
Likelihood of observing these peaks

Likelihood of not having any harmonics in the NP region



$\mathcal{L}_{\text{peak}}(\theta)$ is large

$\mathcal{L}_{\text{non-peak region}}(\theta)$ is small



$\mathcal{L}_{\text{non-peak region}}(\theta)$ is large

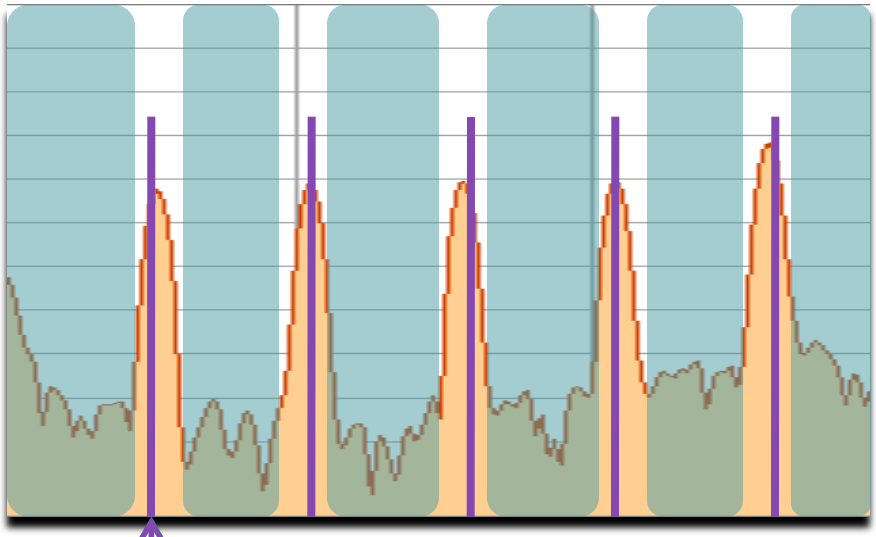
$\mathcal{L}_{\text{peak}}(\theta)$ is small

Likelihood Definition

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{peak}}(\theta) \cdot \mathcal{L}_{\text{non-peak region}}(\theta)$$

Likelihood of observing these peaks

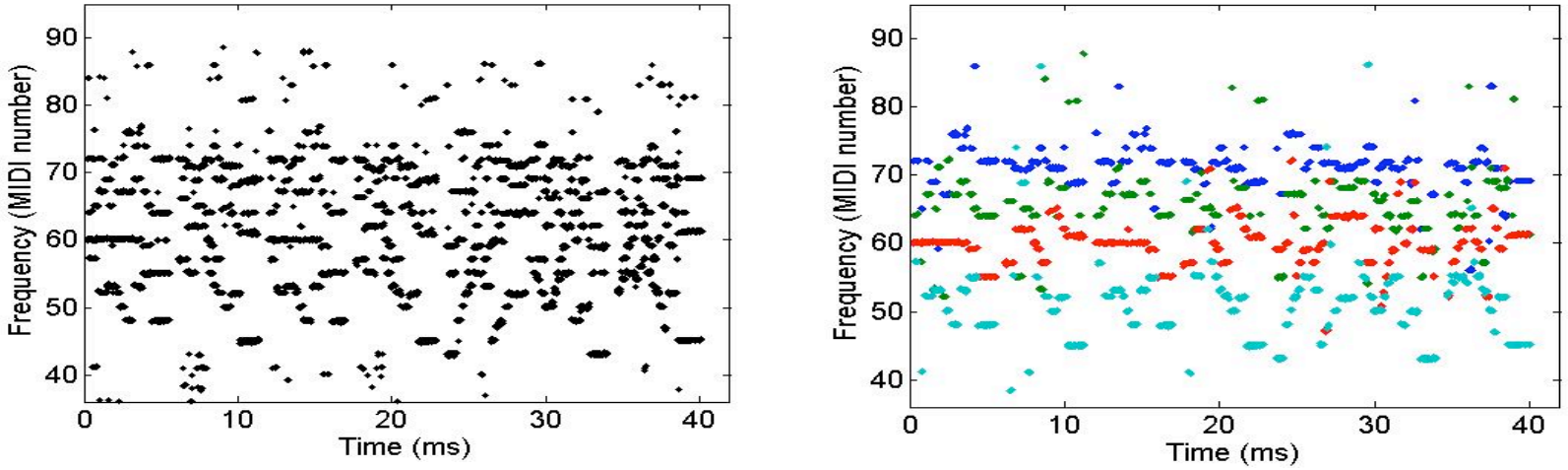
Likelihood of not having any harmonics in the NP region



$\mathcal{L}_{\text{peak}}(\theta)$ is large

$\mathcal{L}_{\text{non-peak region}}(\theta)$ is large

Pitch Trajectory Formation



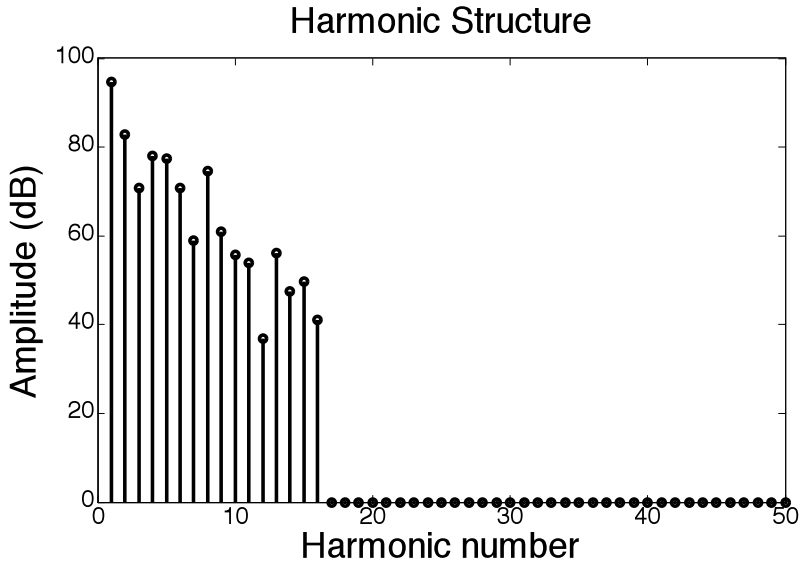
- How to form pitch trajectories ?
  - View it as a constrained clustering problem!
- We use two clustering cues
  - Global timbre consistency
  - Local time-frequency locality

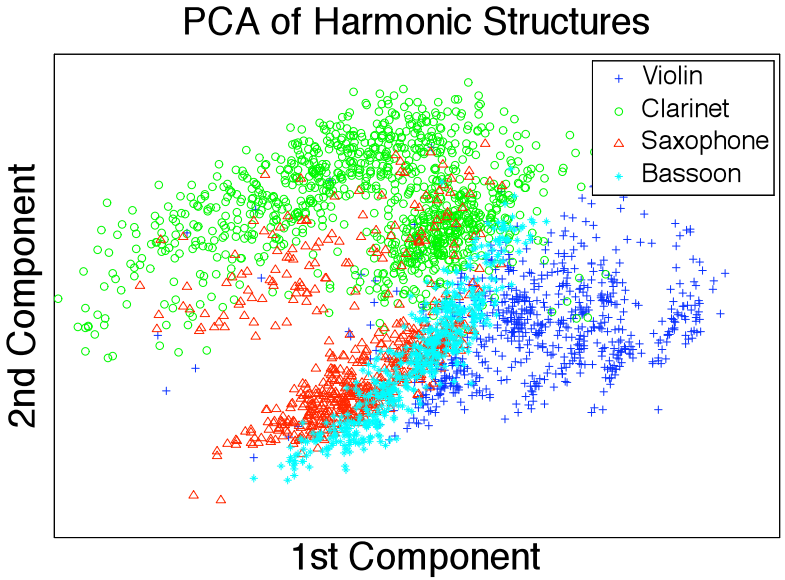
Global Timbre Consistency

- Objective function
  - Minimize intra-cluster distance

$$J = \sum_{k=1}^{K} \sum_{x_i \in T_k} \|\mathbf{x}_i - \mathbf{c}_k\|^2$$

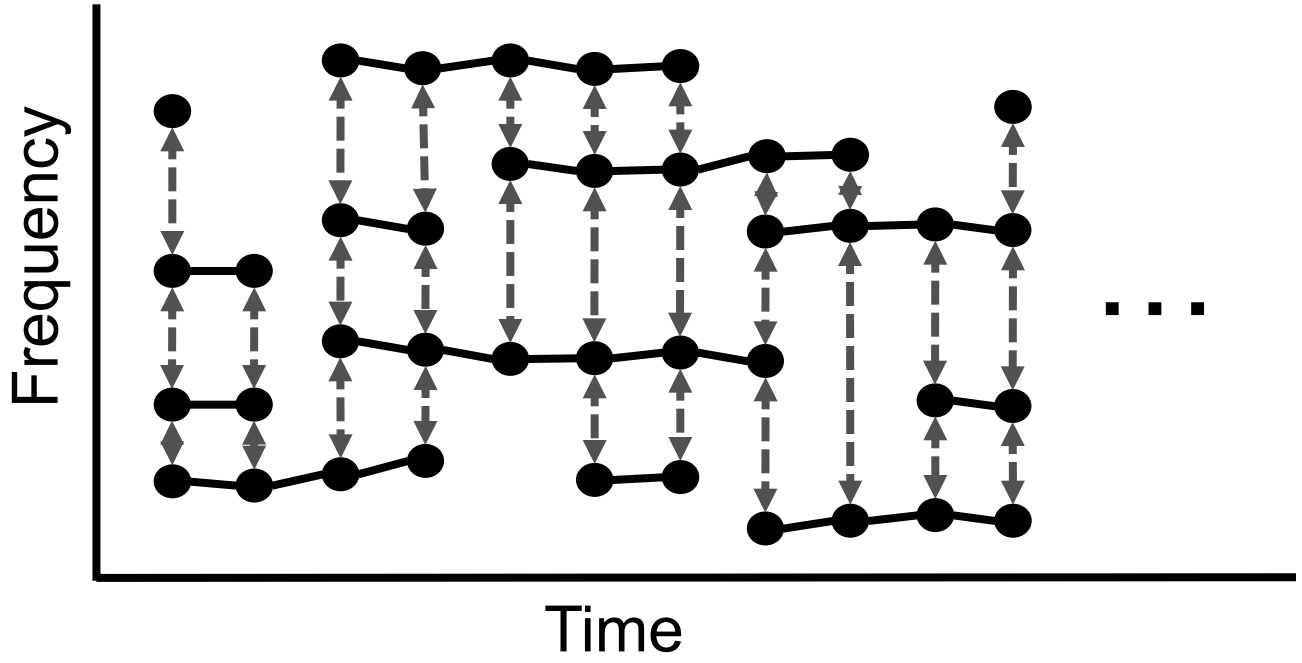
- Harmonic structure feature
  - Normalized relative amplitudes of harmonics





Local Time-frequency Locality

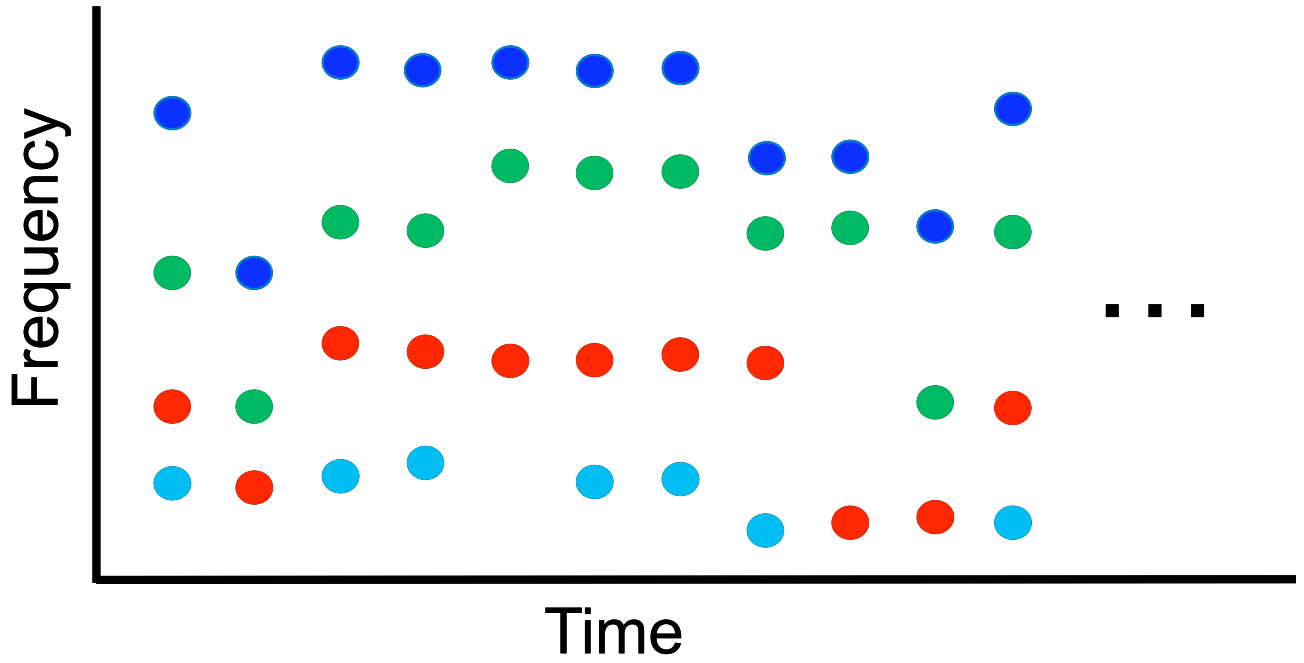
- Constraints
  - Must-link: similar pitches in adjacent frames
  - Cannot-link: simultaneous pitches



- Finding a feasible clustering is NP-hard!

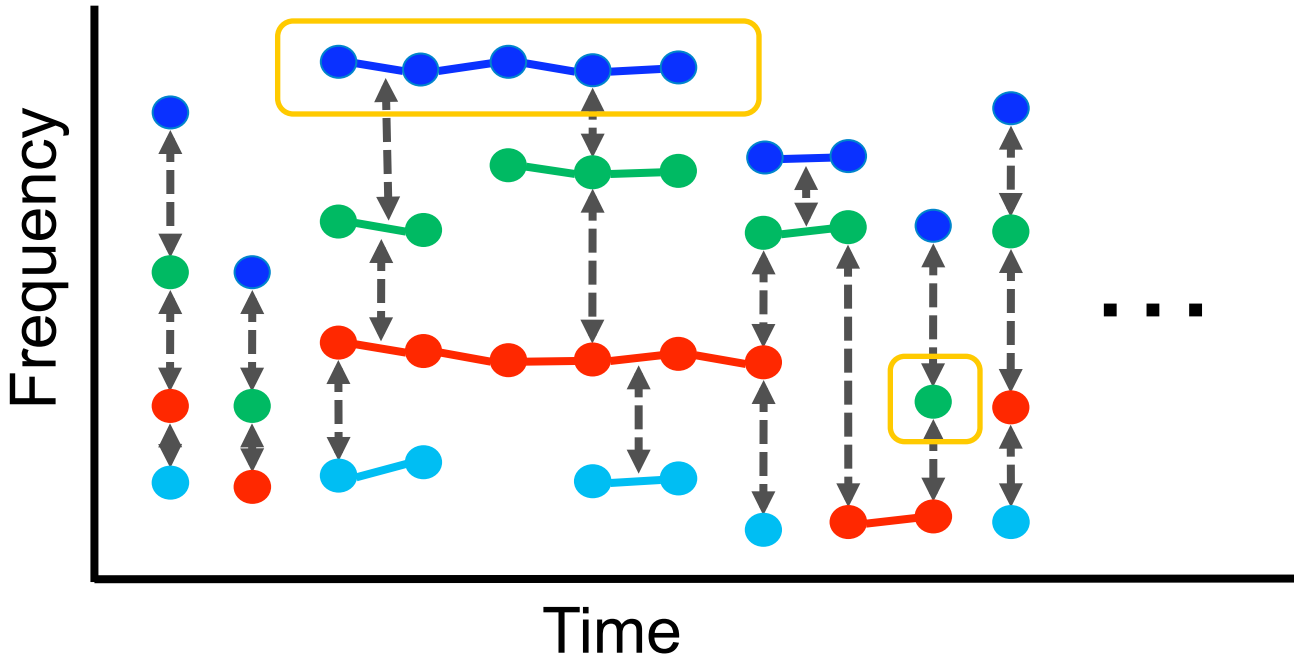
Our Constrained Clustering Process

- 1) Find an initial clustering
  - Labeling pitches according to pitch order in each frame: First, second, third, fourth



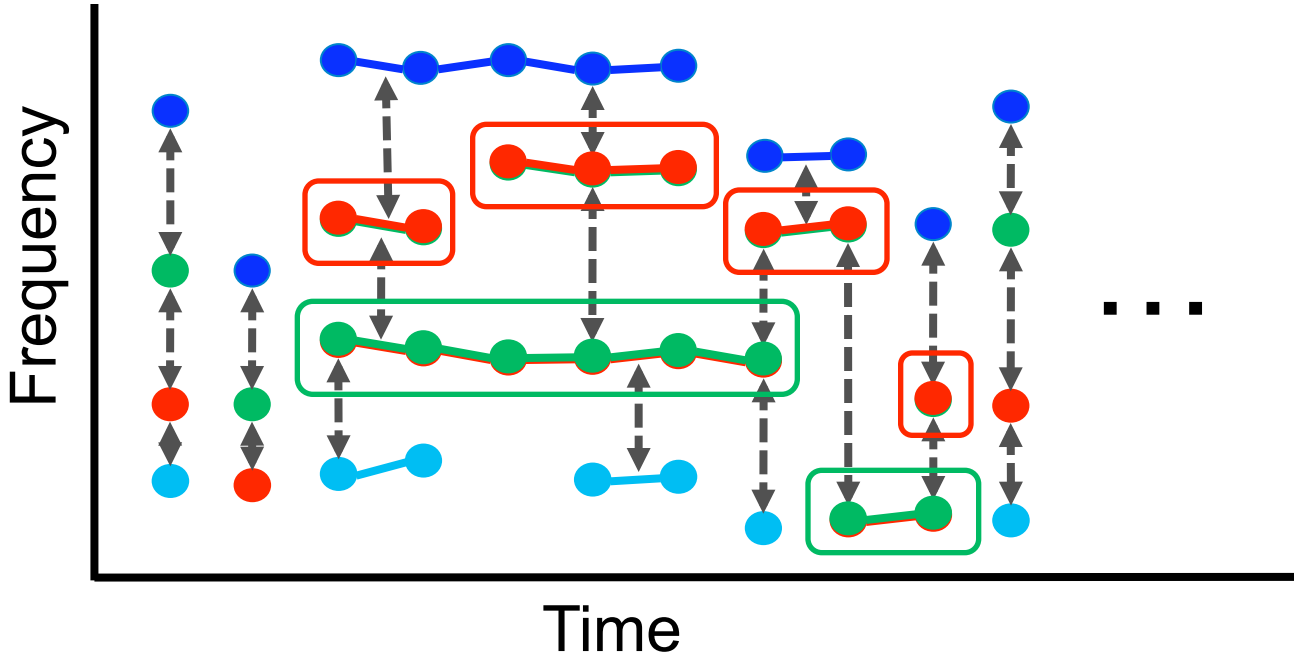
Our Constrained Clustering Process

- 2) Define constraints
  - Must-link: similar pitches in adjacent frames and the same initial cluster: Notelet
  - Cannot-link: simultaneous notelets



Our Constrained Clustering Process

- 3) Update clusters to minimize objective function
  - Swap set: A set of notelets in two clusters connected by cannot-links
  - Swap notelets in a swap set between clusters if it reduces objective function
  - Iteratively traverse all the swap sets



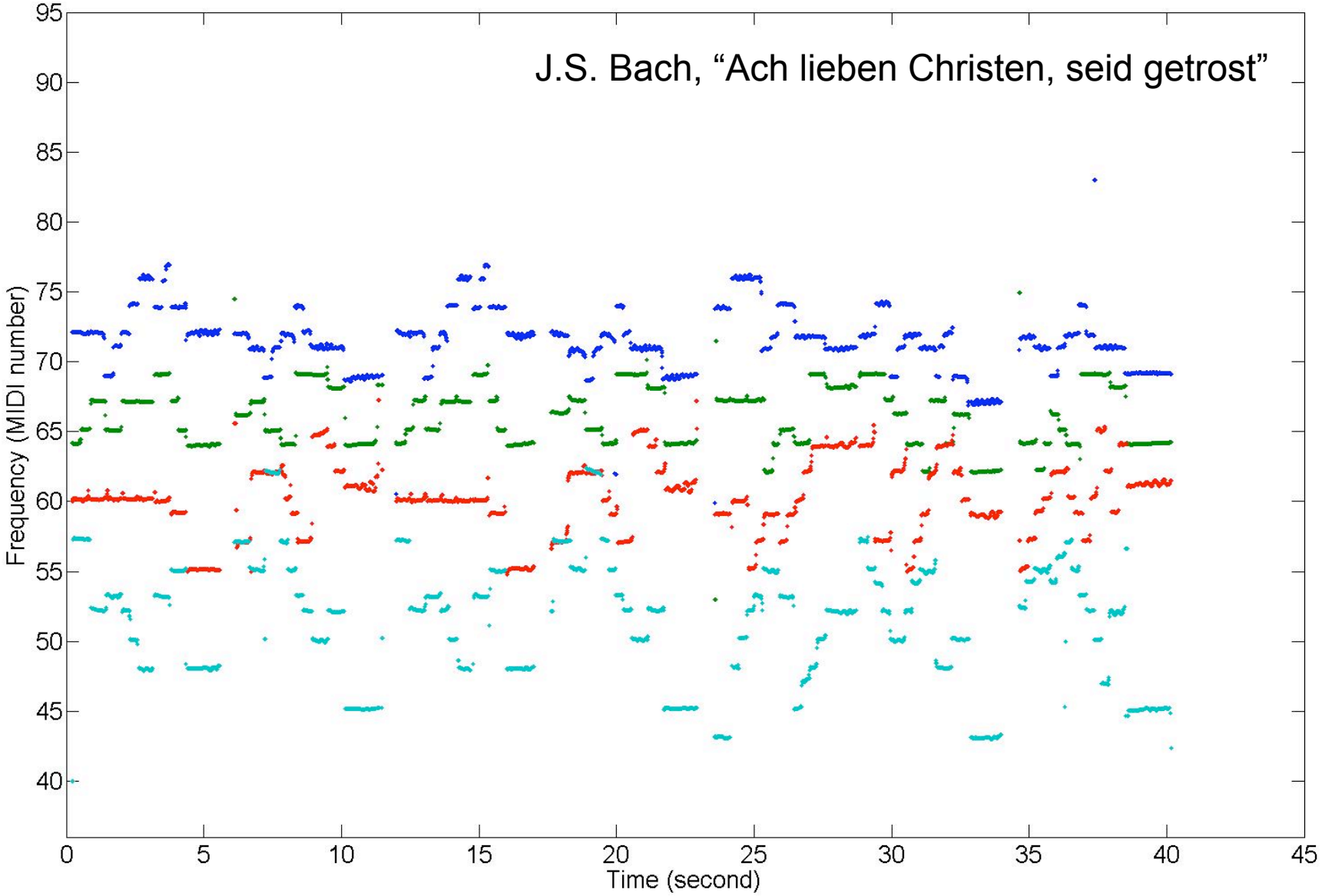
Data Set

- Data set
  - 10 J.S. Bach chorales (quartets, played by violin, clarinet, saxophone and bassoon)
  - Each instrument is recorded individually, then mixed
- Ground-truth pitch trajectories
  - Use YIN on monophonic tracks before mixing

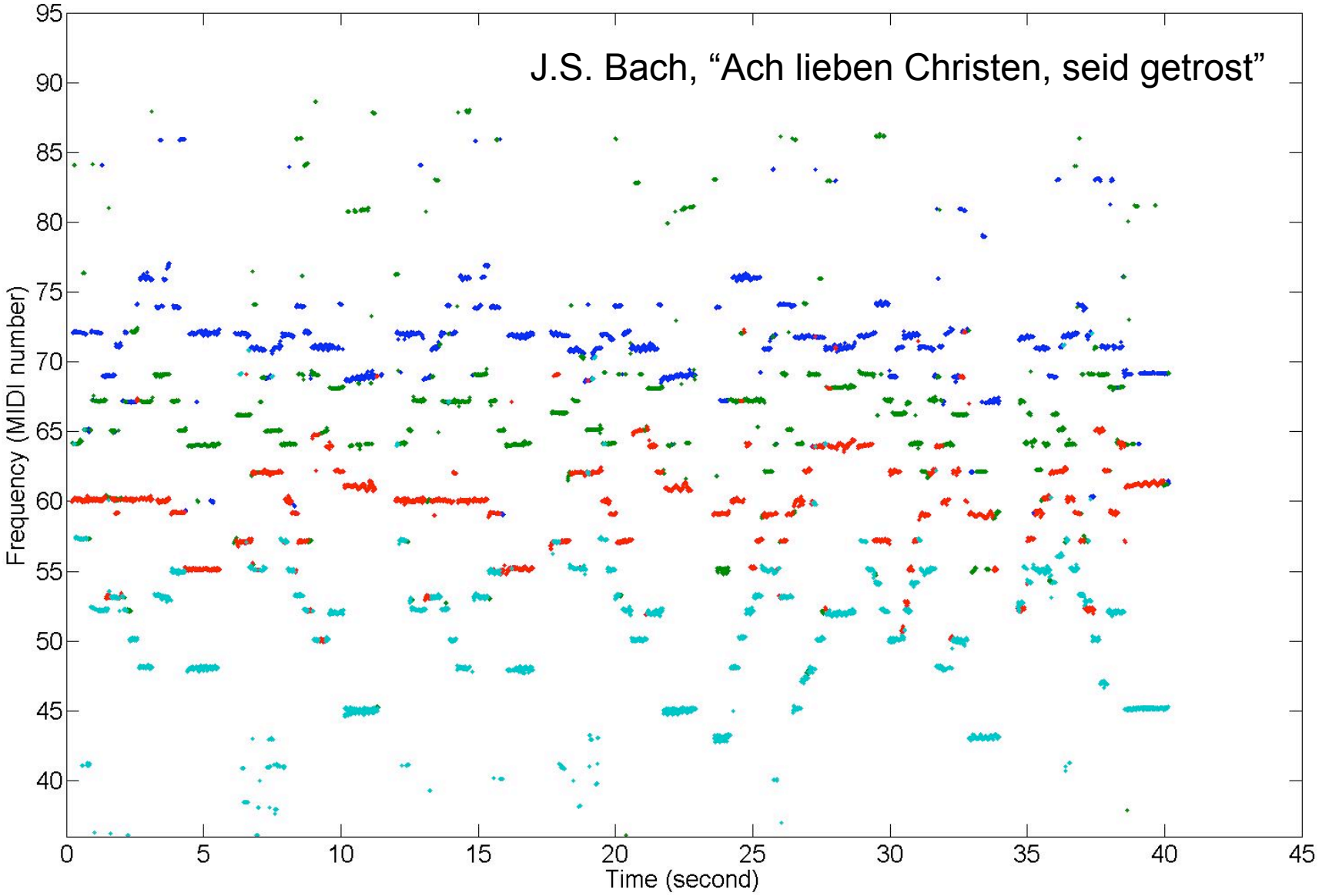
Experimental Results

Mean +- Std		Precision (%)	Recall (%)
How many pitches are correctly estimated?	Klapuri, ISMIR2006	87.2 +- 2.0	66.2 +- 3.4
→	**Ours**	**88.6 +- 1.7**	**77.0 +- 3.5**
How many pitches are correctly estimated and put into the correct trajectory?	Chance	Approx 0.0	Approx 0.0
→	**Ours**	**76.9 +- 11.0**	**67.1 +- 11.9**
How many notes are correctly estimated?	Chance	Approx 0.0	Approx 0.0
→	**Ours**	**46.0 +- 5.5**	**54.3 +- 5.5**

Ground Truth Pitch Trajectories



Our System's Output



Conclusion

- Our multi-pitch tracking system
  - Multi-pitch estimation in single frame
    - Estimate F0s by modeling peaks and the non-peak region
    - Estimate polyphony, refine F0s estimates
  - Pitch trajectory formation
    - Constrained clustering
      - Objective: timbre (harmonic structure) consistency
      - Constraints: local time-frequency locality of pitches
    - A clustering algorithm by swapping labels
- Results on music recordings are promising

Thanks you!
Q & A

Possible Questions

- How much does our constrained clustering algorithm improve from the initial pitch trajectory (label pitches by pitch order)?