Topic 10

Multi-pitch Analysis

What is pitch?

- "Common elements of music are **pitch**, rhythm, dynamics, and the sonic qualities of timbre and texture."

  ---- Wikipedia

- An auditory perceptual attribute in terms of which sounds may be ordered from low to high.

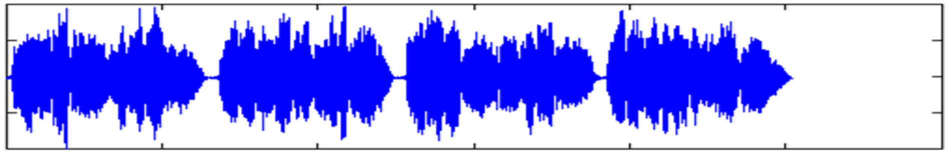
- For (quasi) harmonic sound e.g. a flute note, it is well defined by the Fundamental Frequency (F0).

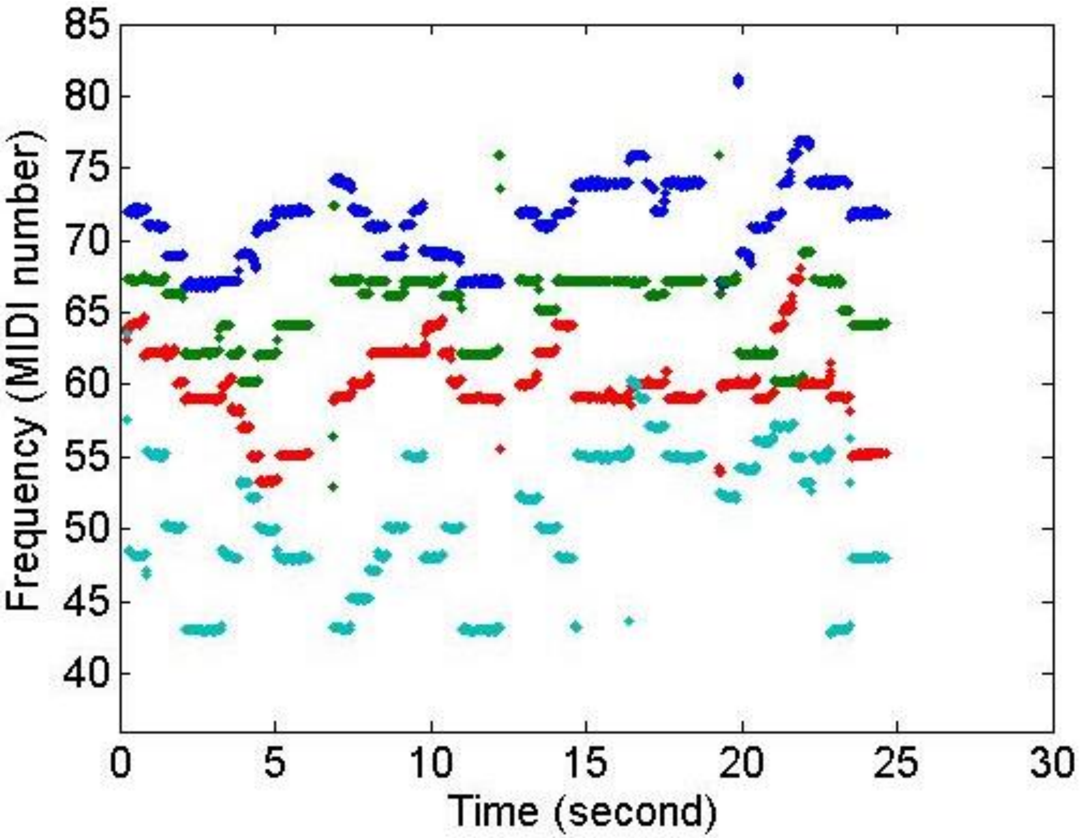
  Oboe C4 | Oboe G4 | Clarinet C4

- A mixture of (quasi) harmonic sounds has multiple pitches (F0s).

Multi-pitch Analysis of Polyphonic Music

- Given polyphonic music played by several harmonic instruments





- Estimate a pitch trajectory for each instrument

Why is it important?

- A fundamental problem in computer audition for harmonic sounds
- Many potential applications
  - Automatic music transcription
  - Harmonic source separation
  - Melody-based music search
  - Chord recognition
  - Music education
  - ……

How difficult is it?

- Let’s do a test!
  - Q1: How many pitches are there?
  - Q2: What are their pitches?
  - Q3: Can you find a pitch in Chord 1 and a pitch in Chord 2 that are played by the same instrument?

Chord 1	Chord 2
2	3
C4/G4	C4/F4/A4
Clarinet G4 Horn C4	Clarinet A4 Viola F4 Horn C4

We humans are amazing!

- "In Rome, he (14 years old) heard Gregorio Allegri's *Miserere* **once** in performance in the Sistine Chapel. He wrote it out **entirely from memory**, only returning to correct **minor errors**…"

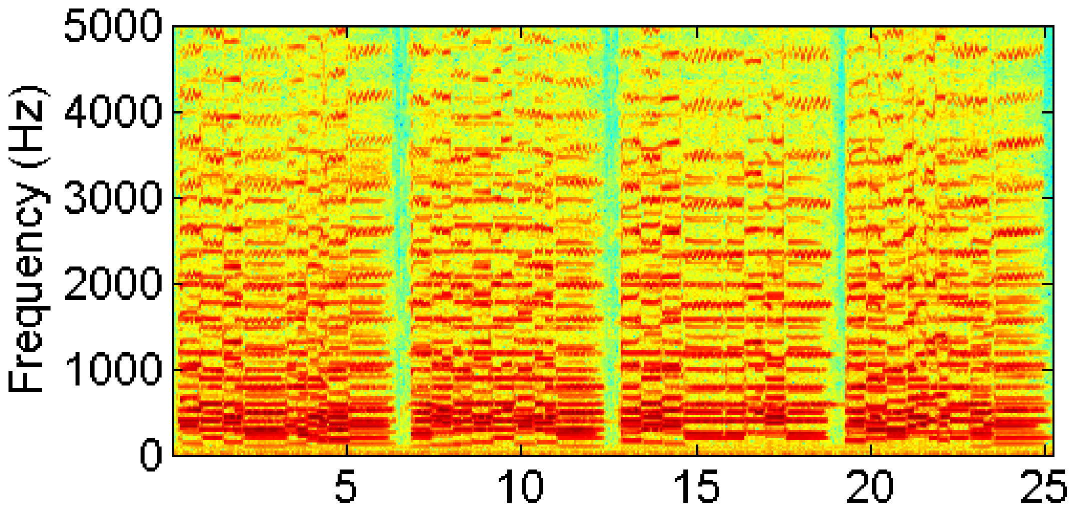
  -- Gutman, Robert (2000). *Mozart: A Cultural Biography*

Wolfgang Amadeus Mozart

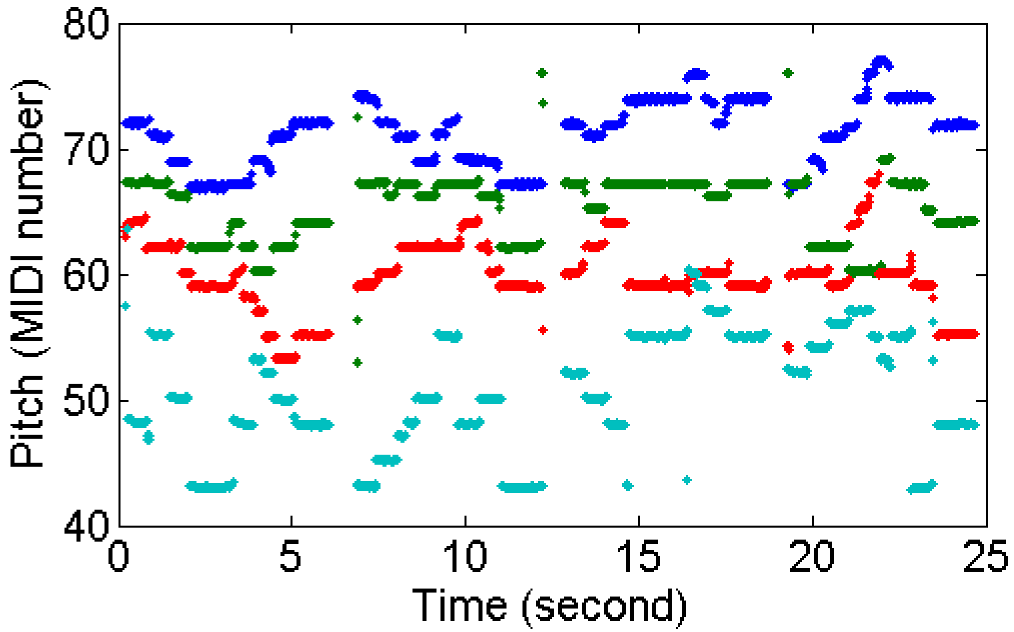
- Can we make computers compete with Mozart??

Our Task

Spectrogram



Ground-truth pitch trajectories



Subtasks in Multi-pitch Analysis

Three levels according to MIREX:

- Level 1: Multi-pitch Estimation (MPE)
  - Estimate pitches and polyphony in each time frame
- Level 2: Note Tracking
  - Track pitches within a note
- Level 3: Streaming (timbre tracking)
  - Estimate a pitch trajectory for each source (instrument) across multiple notes

Recent Methods

- Level 1: Multi-pitch Estimation
  - Klapuri'03, Goto'04, Davy'06, Klapuri'06, Yeh'05, Emiya'07, Pertusa'08, Duan'10, etc.

- Level 2: Note Tracking
  - Ryynanen'05, Kameoka'07, Poliner'07, Lagrange'07, Chang'08, Benetos'11, Cogliati'16, Ewert'17, Hawthorne'18, etc.

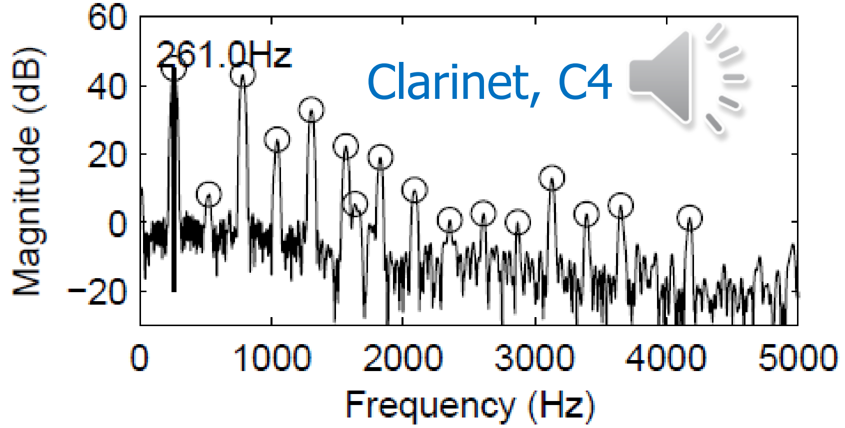
- Level 3: Streaming (timbre tracking)
  - Vincent'06, Bay'12, Duan'14

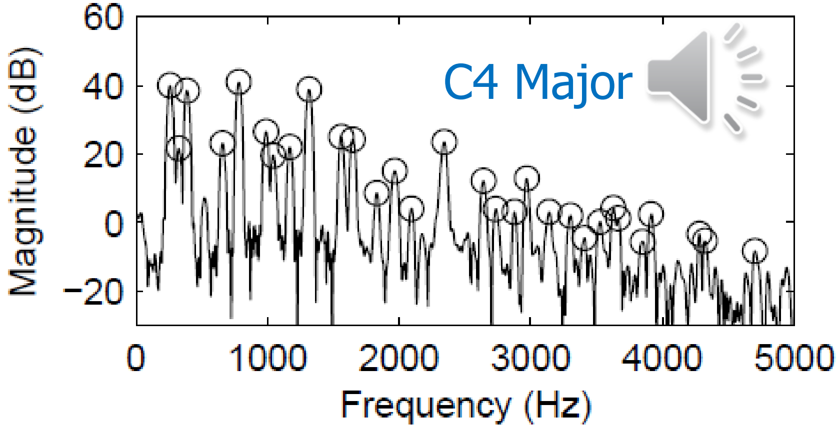
Level 1: Multi-pitch Estimation

Estimate pitches in each single frame

Multi-pitch Estimation (MPE)

- Why difficult?





  - Overlapping harmonics
    - C4 (46.7%), E4 (33.3%), G4 (60%)
  - How to associate the 28 significant peaks to sources?
  - Instantaneous polyphony estimation
  - Large hypothesis space

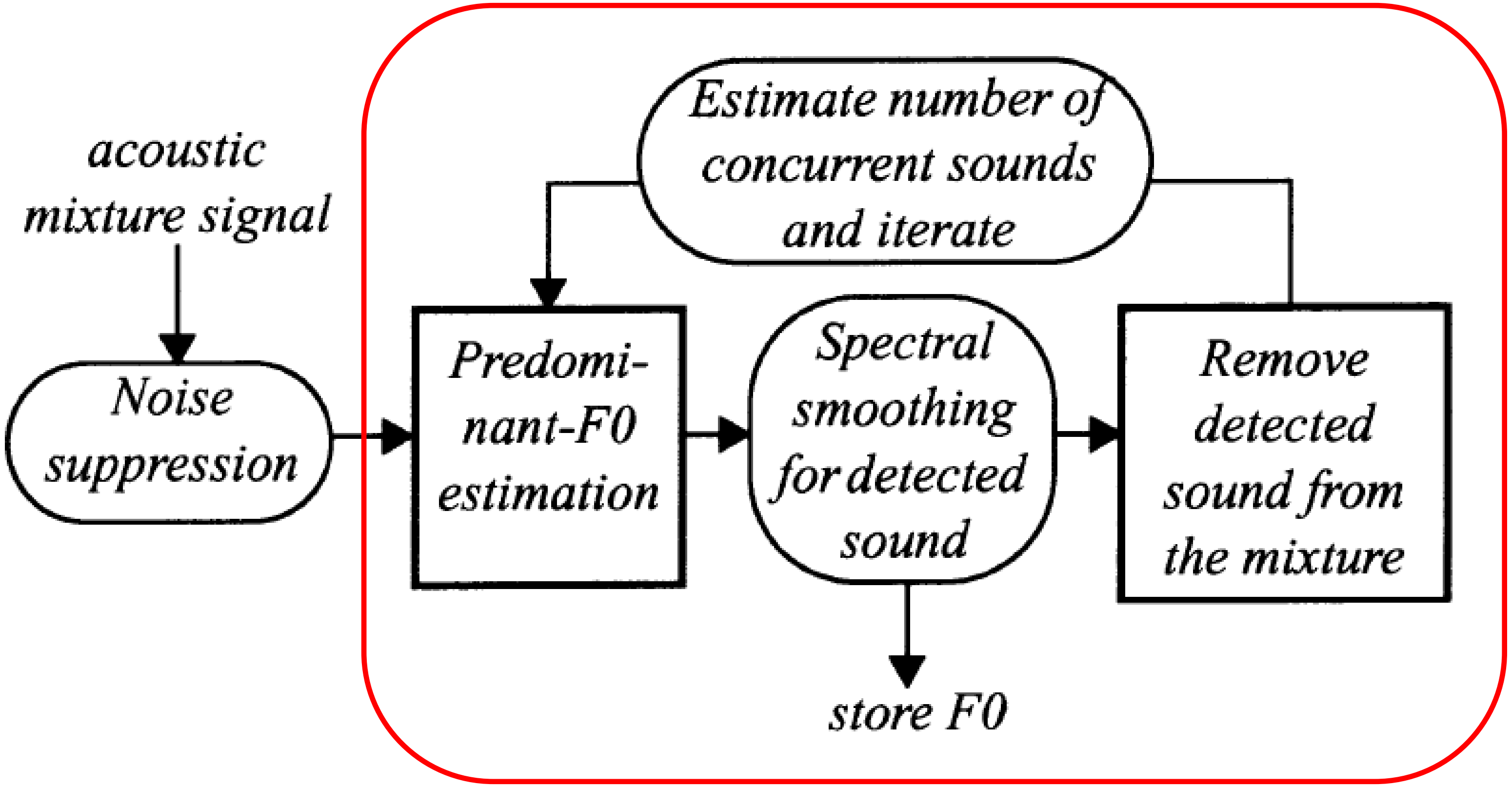
Two Methods at Level 1

- Iterative spectral subtraction
  - [Klapuri, 2003]

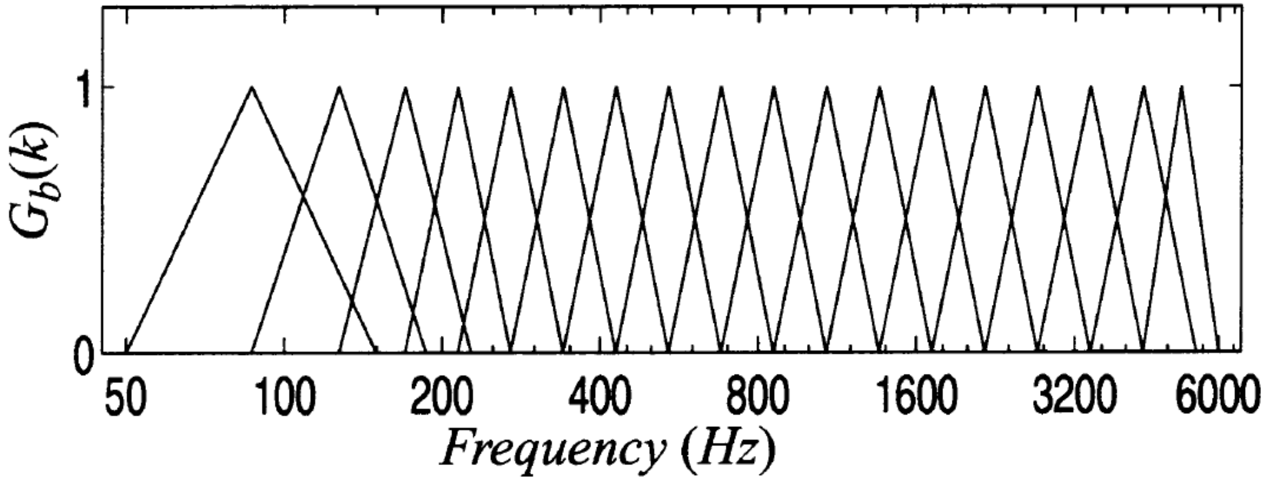
- Probabilistic modeling of peaks and non-peak regions
  - [Duan et al., 2010]

Iterative Spectral Subtraction

[Klapuri, 2003]



Bandwise F0 Estimation



$$Z_b(k) = G_b(k)Z(k)$$

magnitude spectrum in Band b

original magnitude spectrum (noise reduced)

Bandwise F0 Estimation

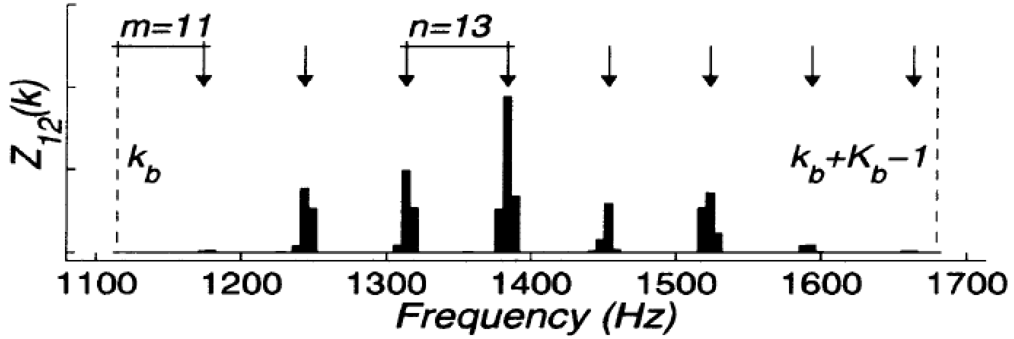
$$L_b(n) = \max_{m \in \boldsymbol{M}} \left\{ c(m,n) \sum_{j=0}^{J(m,n)-1} Z_b(k_b + m + nj) \right\}$$

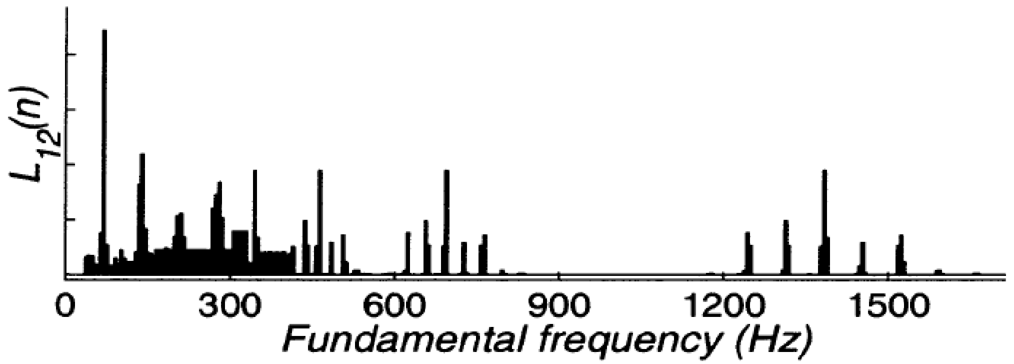
Weight of F0 hyp, n

Normalization factor

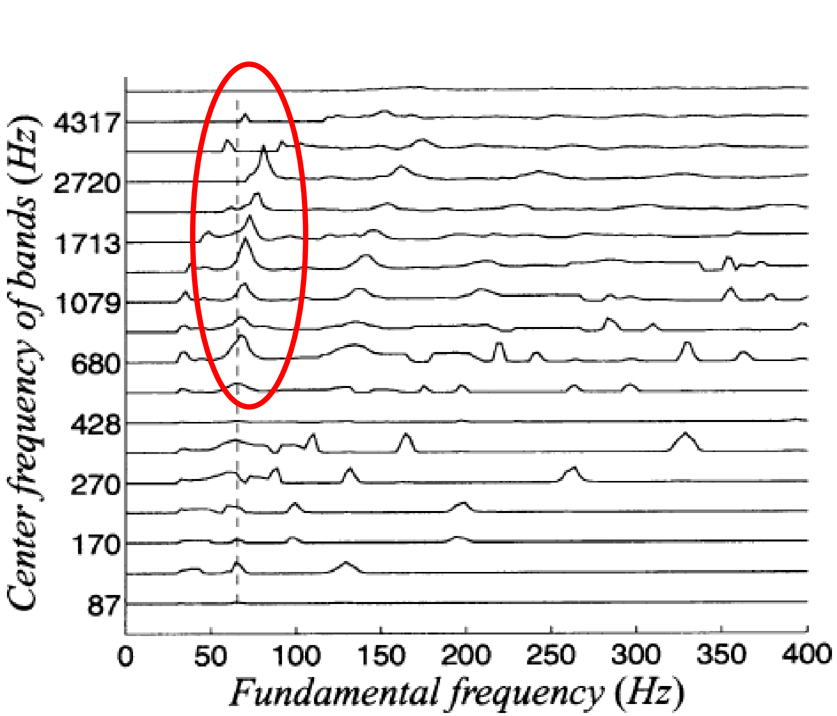
# of partials

Freq. offset

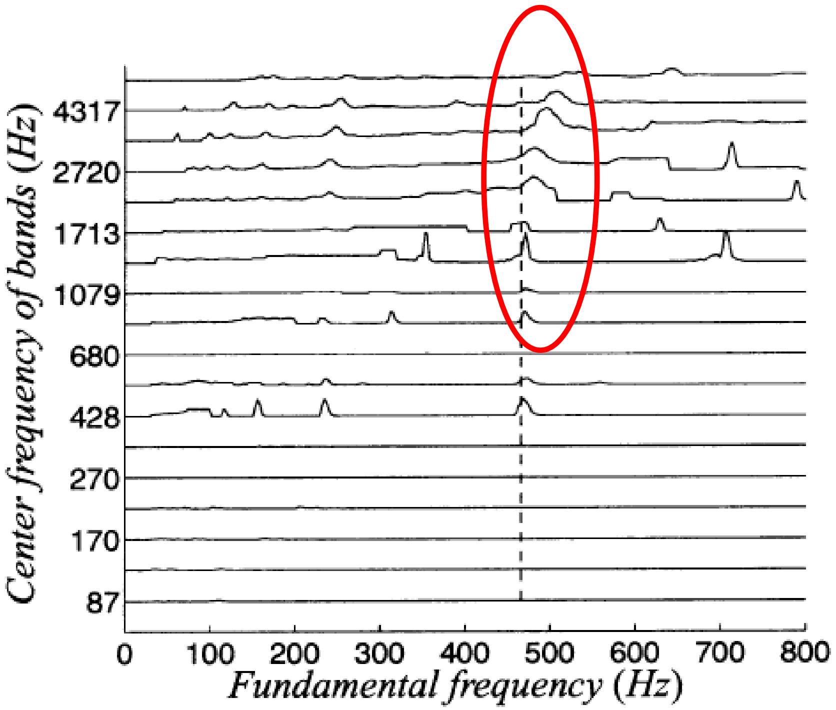




Integrate Weights Across Subbands



Piano note (65Hz)



Piano note (470Hz)

- Inharmonicity of higher harmonics should be considered

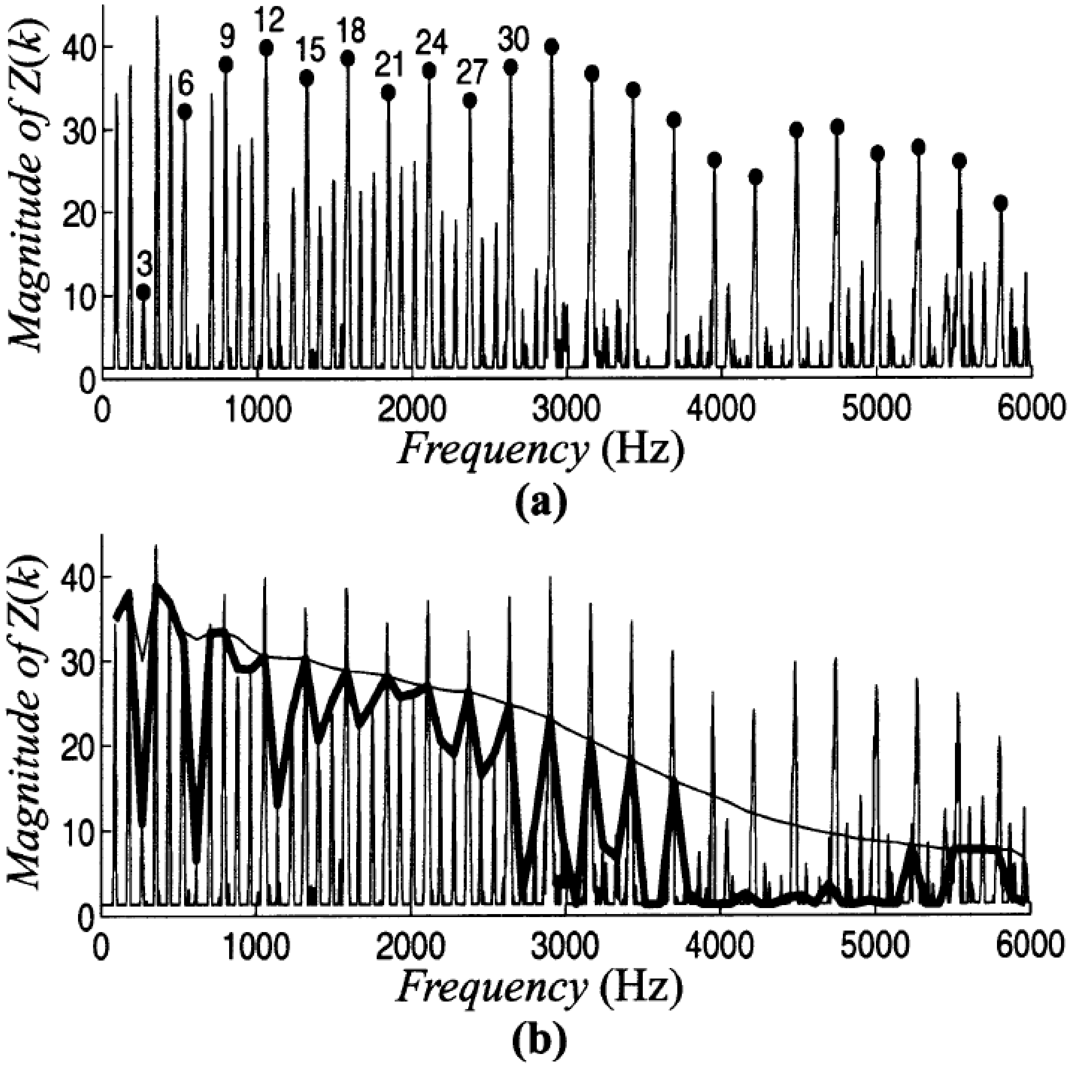
$$f_h = hF\sqrt{1 + (h^2 - 1)\beta}$$

Spectral Subtraction

- Given the estimated predominant F0, we can find out all its harmonics and subtract their energy from the mixture spectrum.

- How much energy should we subtract?
  - All?
  - Some harmonics are overlapped by those of other F0s, hence their energy is larger.

Spectral Smoothness

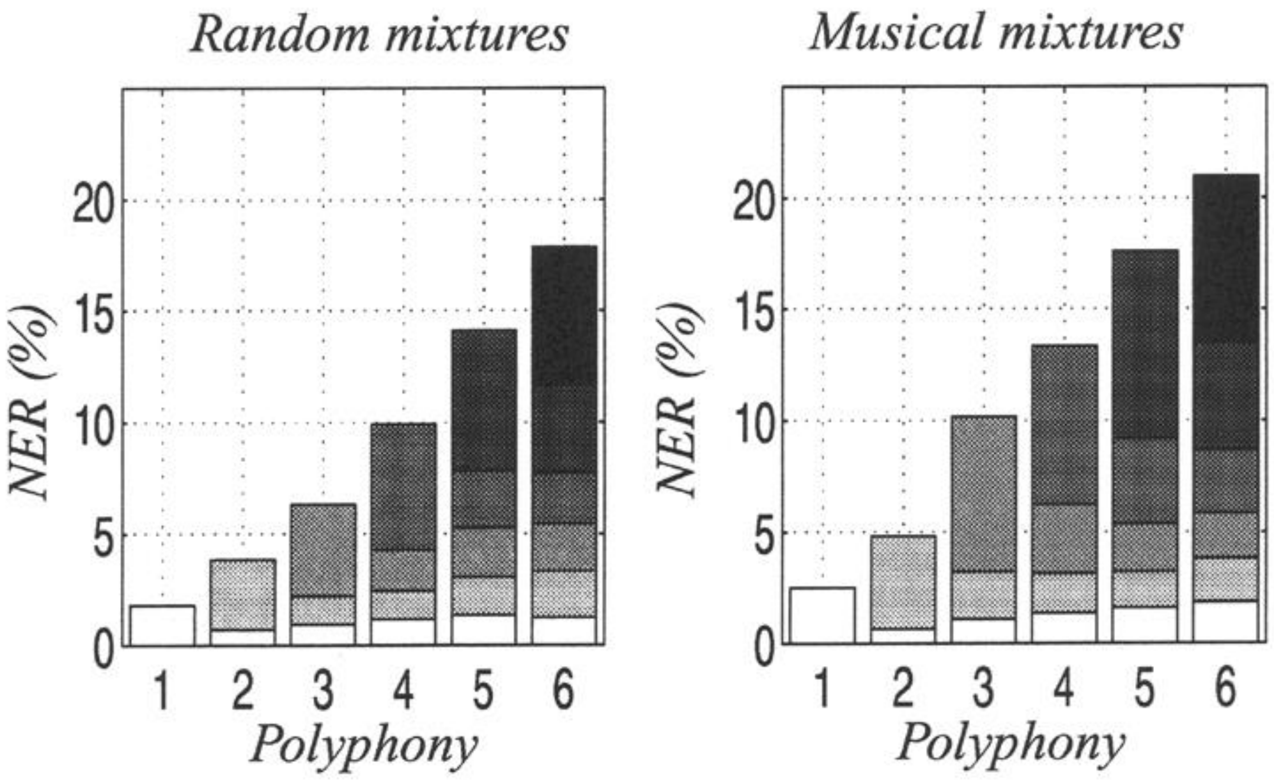


Polyphony Estimation

- I.e., when to stop the iterations?

- Stop if the energy of the harmonics of the estimated predominant F0 is smaller than a threshold.

Error Rate



- More errors in later iterations

Discussions

- Advantages
  - Simple idea
  - Fast algorithm
  - Handles inharmonicity
- Disadvantages
  - Spectra in later iterations are severely corrupted
  - Spectral smoothness is not enough to determine the amount of energy to subtract
- Why bandwise estimation?

Probabilistic Modeling of Peaks

- A maximum likelihood estimation method

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} p(\boldsymbol{O}|\boldsymbol{\theta})$$

[Duan et al., 2010]

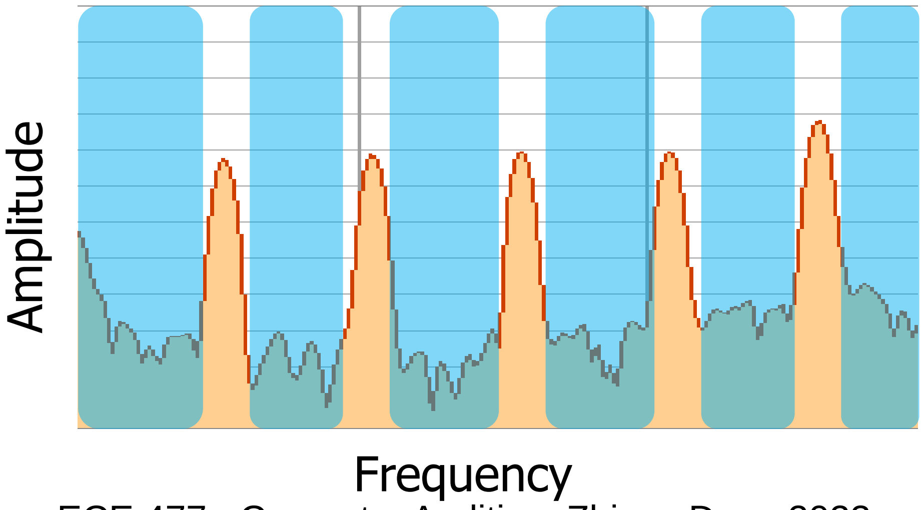
Best pitch estimate (a set of pitches)

Observed power spectrum

Pitch hypothesis, (a set of pitches)

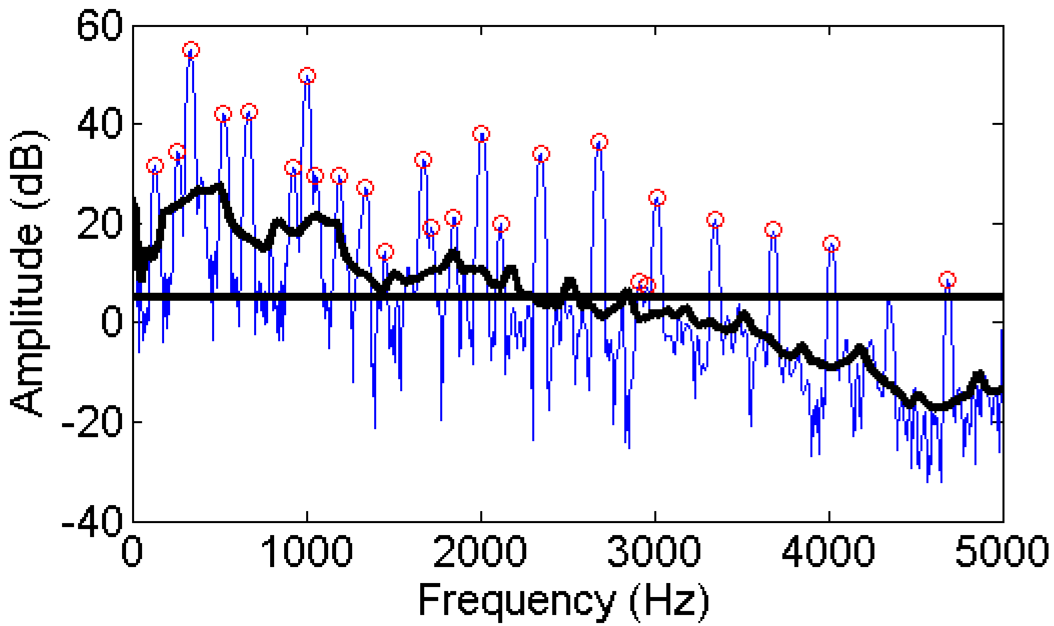
- Spectrum: peaks & the non-peak region

Fourier Transform Power Spectrum:



Peaks / Non-peak Region

- Peaks: ideally correspond to harmonics



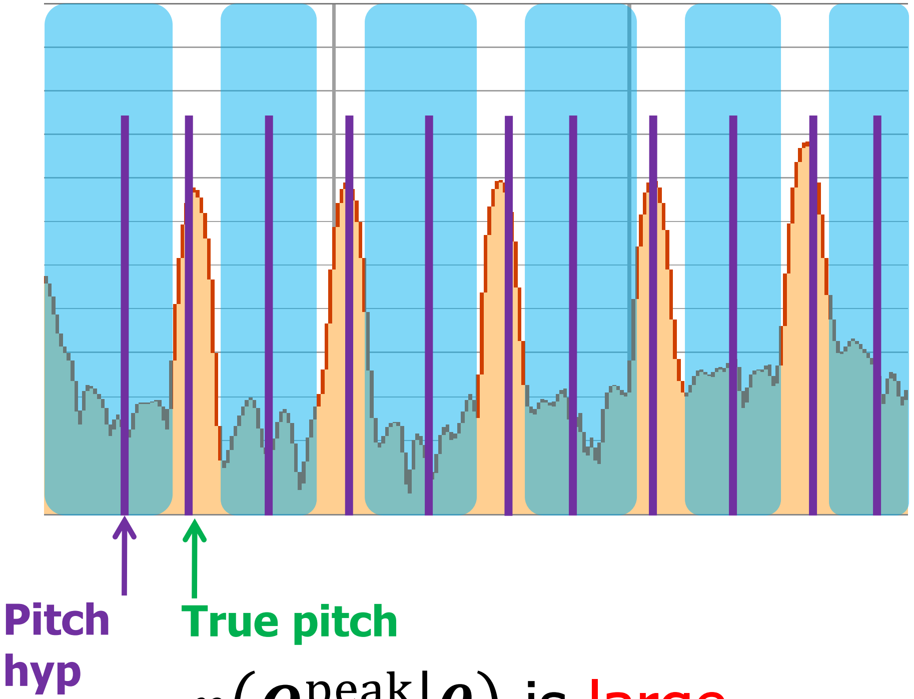
- Non-peak region: frequencies further than a threshold from any peak

Likelihood as Dual Parts

$$p(\boldsymbol{O}|\boldsymbol{\theta}) = p\left(\boldsymbol{O}^{\text{peak}}\middle|\boldsymbol{\theta}\right) \cdot p\left(\boldsymbol{O}^{\text{non-peak}}\middle|\boldsymbol{\theta}\right)$$

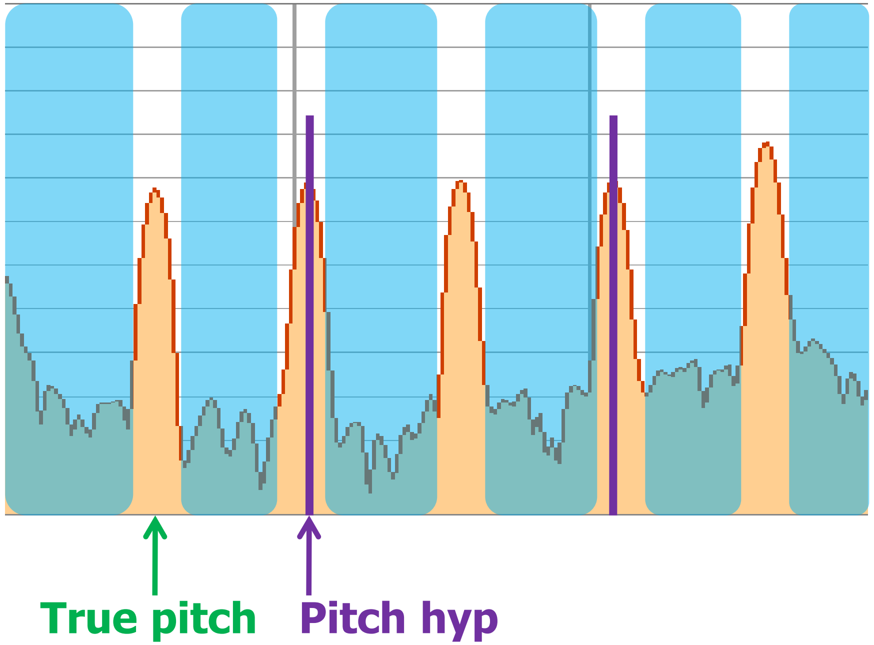
Probability of observing these peaks: $(f_k, a_k), k = 1, \dots, K$.

Probability of not having any harmonics in the non-peak region



$p\left(\boldsymbol{O}^{\text{peak}}\middle|\boldsymbol{\theta}\right)$ is large

$p\left(\boldsymbol{O}^{\text{non-peak}}\middle|\boldsymbol{\theta}\right)$ is small



$p\left(\boldsymbol{O}^{\text{peak}}\middle|\boldsymbol{\theta}\right)$ is small

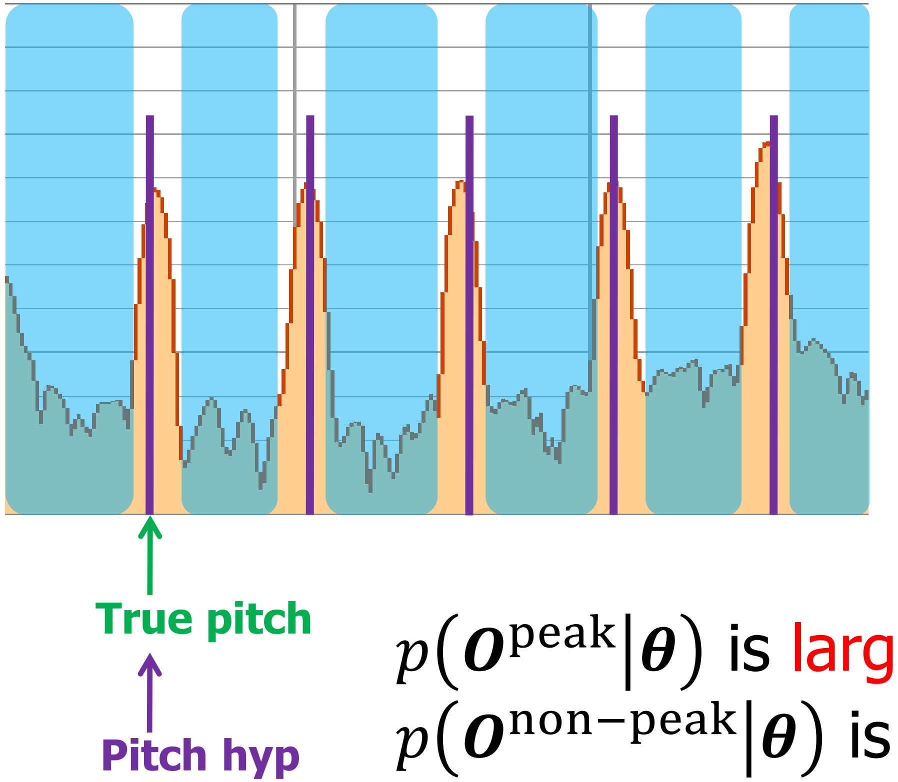
$p\left(\boldsymbol{O}^{\text{non-peak}}\middle|\boldsymbol{\theta}\right)$ is large

Likelihood as Dual Parts

$$p(\boldsymbol{O}|\boldsymbol{\theta}) = p\left(\boldsymbol{O}^{\text{peak}}\middle|\boldsymbol{\theta}\right) \cdot p\left(\boldsymbol{O}^{\text{non-peak}}\middle|\boldsymbol{\theta}\right)$$

Probability of observing these peaks: $(f_k, a_k), k = 1, \ldots, K$.

Probability of not having any harmonics in the non-peak region



$p\left(\boldsymbol{O}^{\text{peak}}\middle|\boldsymbol{\theta}\right)$ is large

$p\left(\boldsymbol{O}^{\text{non-peak}}\middle|\boldsymbol{\theta}\right)$ is large

Likelihood Models

$$p(\boldsymbol{O}^{\text{peak}}|\boldsymbol{\theta}) \approx \prod_{k=1}^{K} p\left(f_k, a_k|\boldsymbol{\theta}\right)$$

Frequency and Amplitude of the k-th peak

Probability of observing these peaks

$$p(\boldsymbol{O}^{\text{non-peak}}|\boldsymbol{\theta}) \approx \prod_{F_0\in\boldsymbol{\theta}} \prod_{\substack{h\in\{1\cdots H\} \\ F_h\in\mathcal{F}_{\text{np}}}} 1 - P\left(e_h = 1|F_0\right)$$

Probability of not having any harmonics in the non-peak region

Freq of the h-th harmonic

The h-th harmonic of F0 exists or not

Learned from training data

Model Training

- For polyphonic music
  - 3000 random chords of polyphony 1 to 6
  - Mixed using note samples from 16 instruments with pitch ranges from C2 (65 Hz) to B6 (1976 Hz)
- For multi-talker speech
  - 500 speech excerpts with 1-3 simultaneous talkers
  - Mixed from single-talker speech
- Obtained ground-truth pitches before mixing

Greedy Search Algorithm

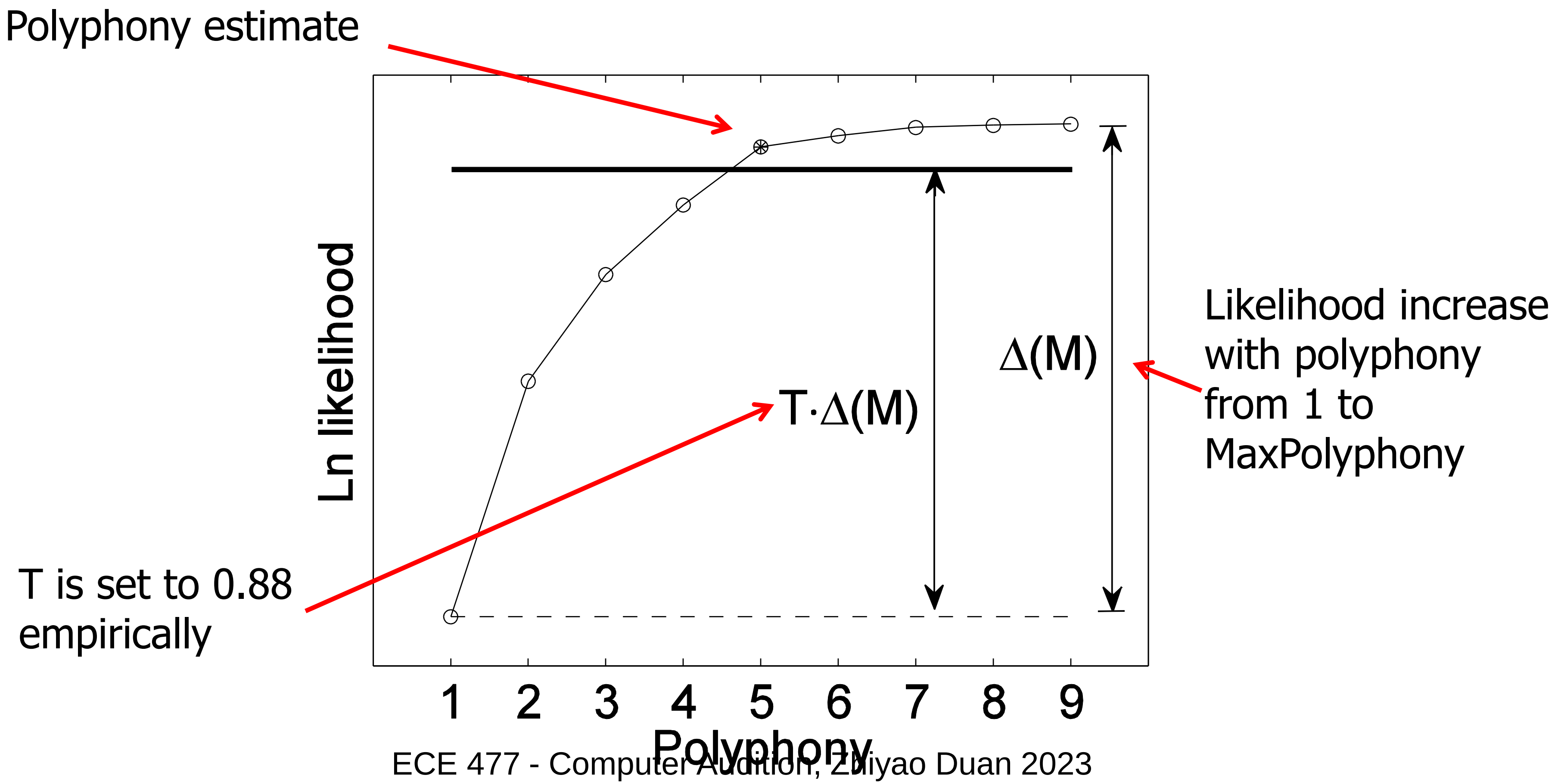
$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} p(\boldsymbol{O}|\boldsymbol{\theta})$$

- Parameter space is too big for exhaustive search
- Greedy search algorithm
  - Initialize $\boldsymbol{\theta} = \emptyset$
  - For $i = 1$ to $MaxPolyphony$
    - Add a pitch to $\boldsymbol{\theta}$, s.t. likelihood increases
  - End
  - Estimate polyphony N
  - Return the first N pitches of $\boldsymbol{\theta}$

Polyphony Estimation

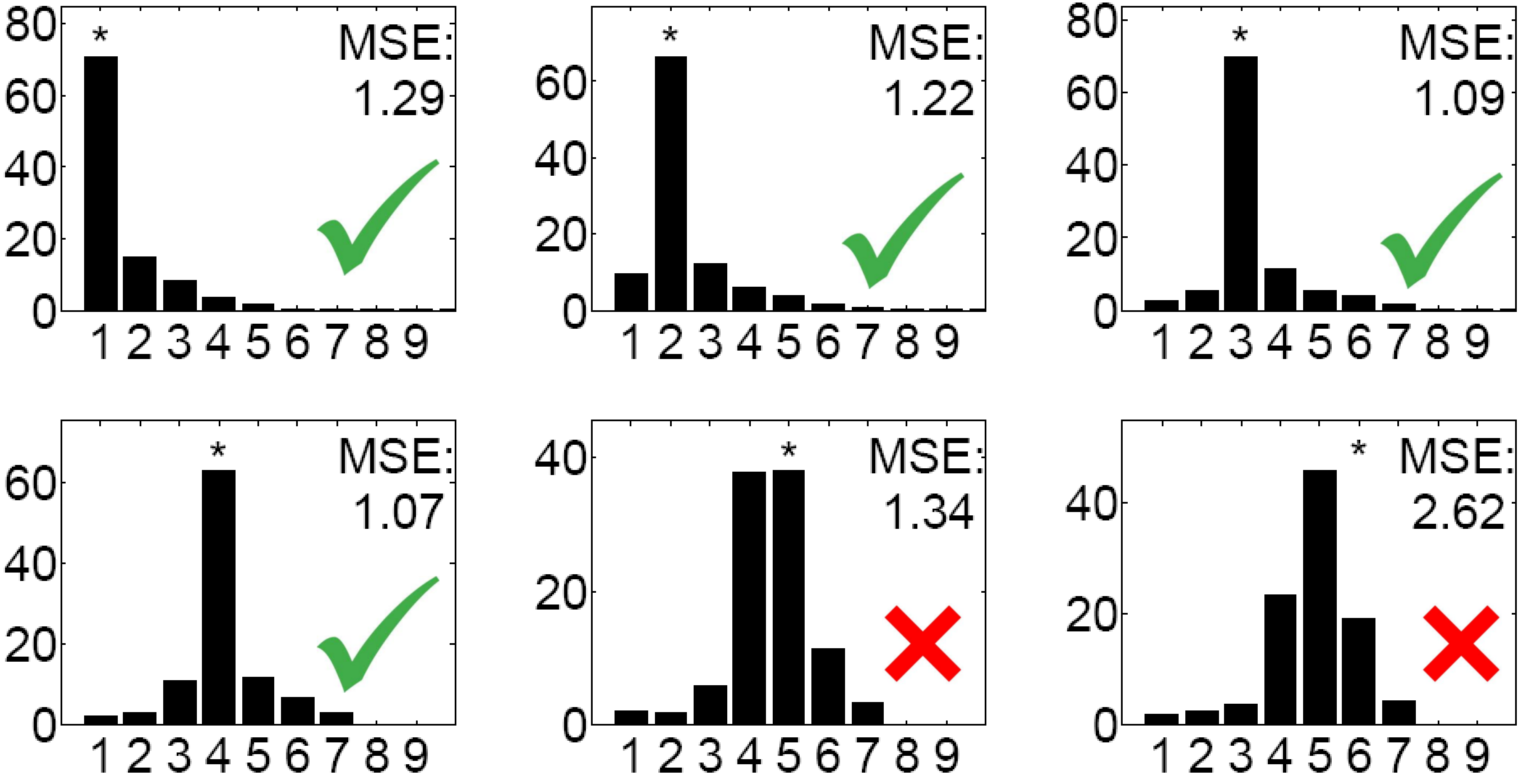
- Likelihood increases with estimated polyphony

$$\mathcal{L}(\hat{\boldsymbol{\theta}}^{n}) \leq \mathcal{L}(\hat{\boldsymbol{\theta}}^{n+1})$$

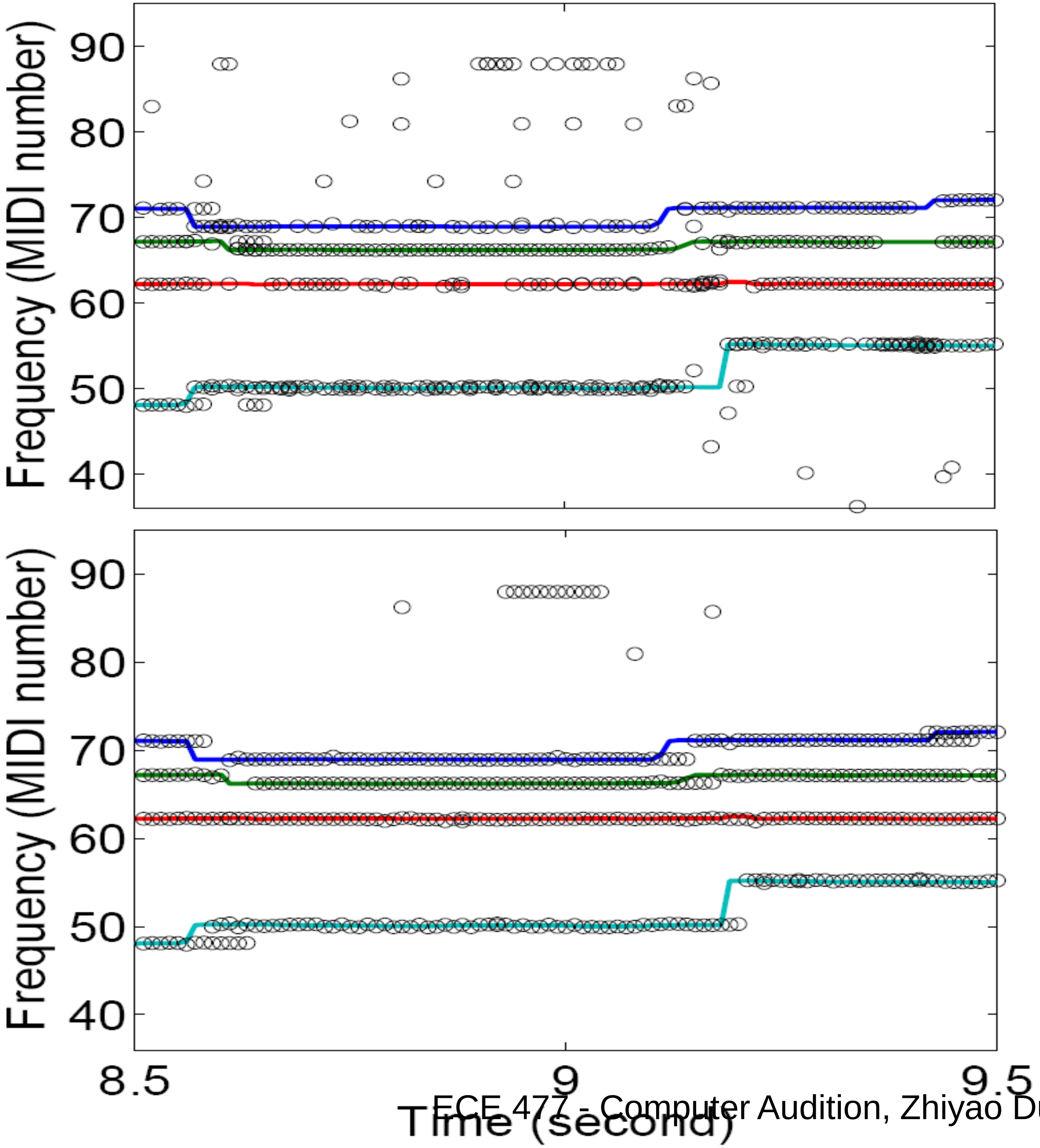


Experiments – Polyphony Estimation

- 6000 musical chords mixed using notes unseen in training data (1000 for each polyphony)



Post Processing



- Estimation in each single frame is not robust
  - Insertion, deletion and substitution errors
- Refine estimates using neighboring frames
  - Only keep consistent estimates

Discussions

- Advantages
  - Model parameters can be learned from training data

- Disadvantages
  - Assumes conditional independence of peak amplitudes, given F0s
  - Doesn't consider the relation between peak amplitudes, e.g., spectral smoothness

Level 2: Note Tracking

Estimate a pitch trajectory for each note

Two Methods at Level 2

- Probabilistic modeling of the spectral-temporal content a note of a source
  - [Kameoka, et al., 2007]

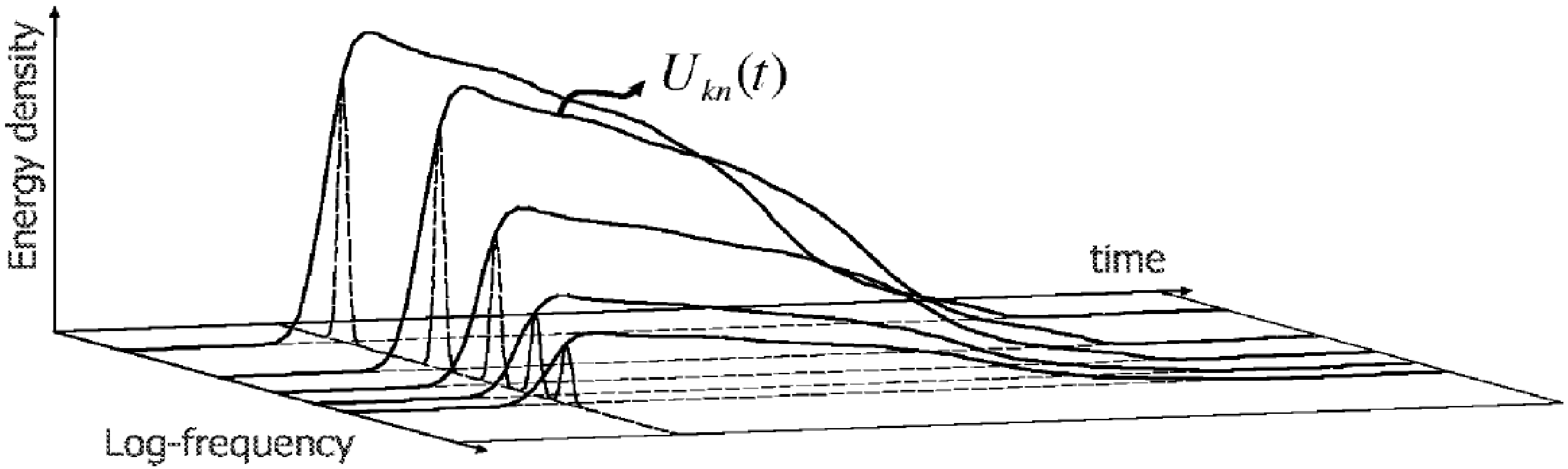
- Classification-based piano note transcription
  - [Poliner & Ellis, 2007]
  - [Hawthorne, et al., 2018]

Harmonic Temporal Structured Clustering (HTC)

[Kameoka et al, 2007]

- Jointly estimates pitch, intensity, onset, duration of notes.
- Detailed parametric model for the spectral content of a note of a source
- Approximating the spectrogram with superimposed HTC source models

HTC Source Model



Relative energy of n-th harmonic

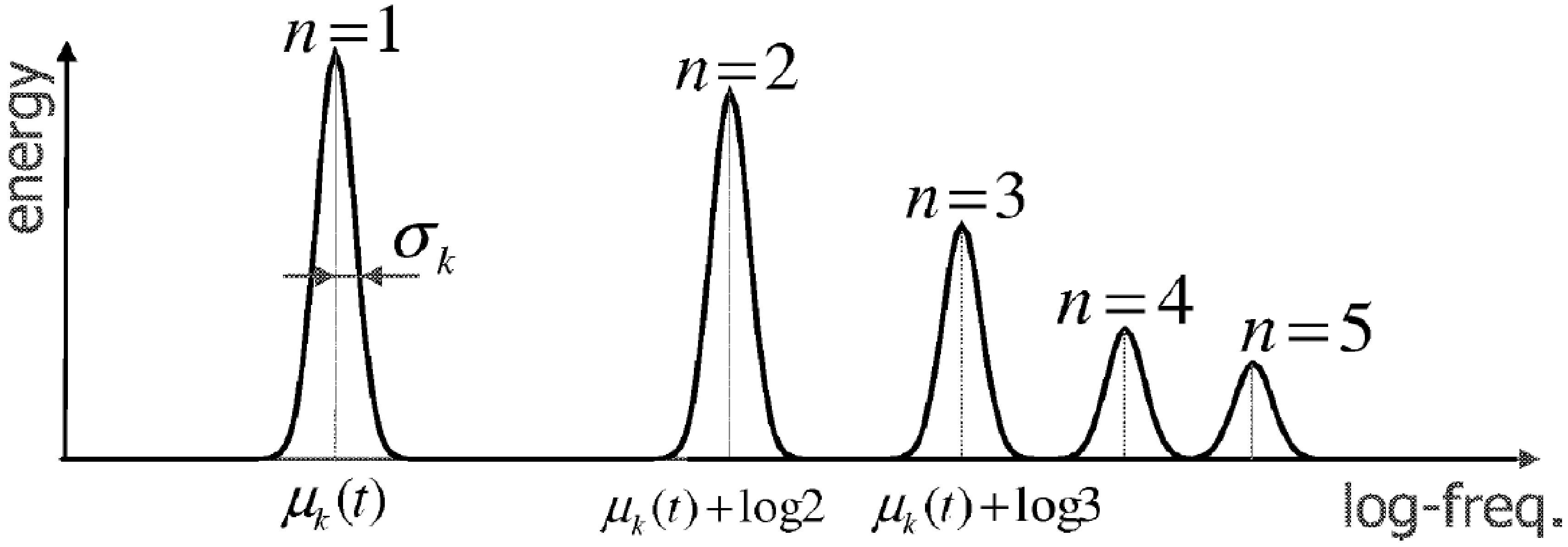
Harmonic envelope over time

$$q_k(x,t;\mathbf{\Theta}) = w_k \sum_{n=1}^{N} \frac{v_{k,n}U_{k,n}(t)}{\sqrt{2\pi}\sigma_k} e^{-(x-\mu_k(t)-\log n)^2/2{\sigma_k}^2}$$

Total energy of the source

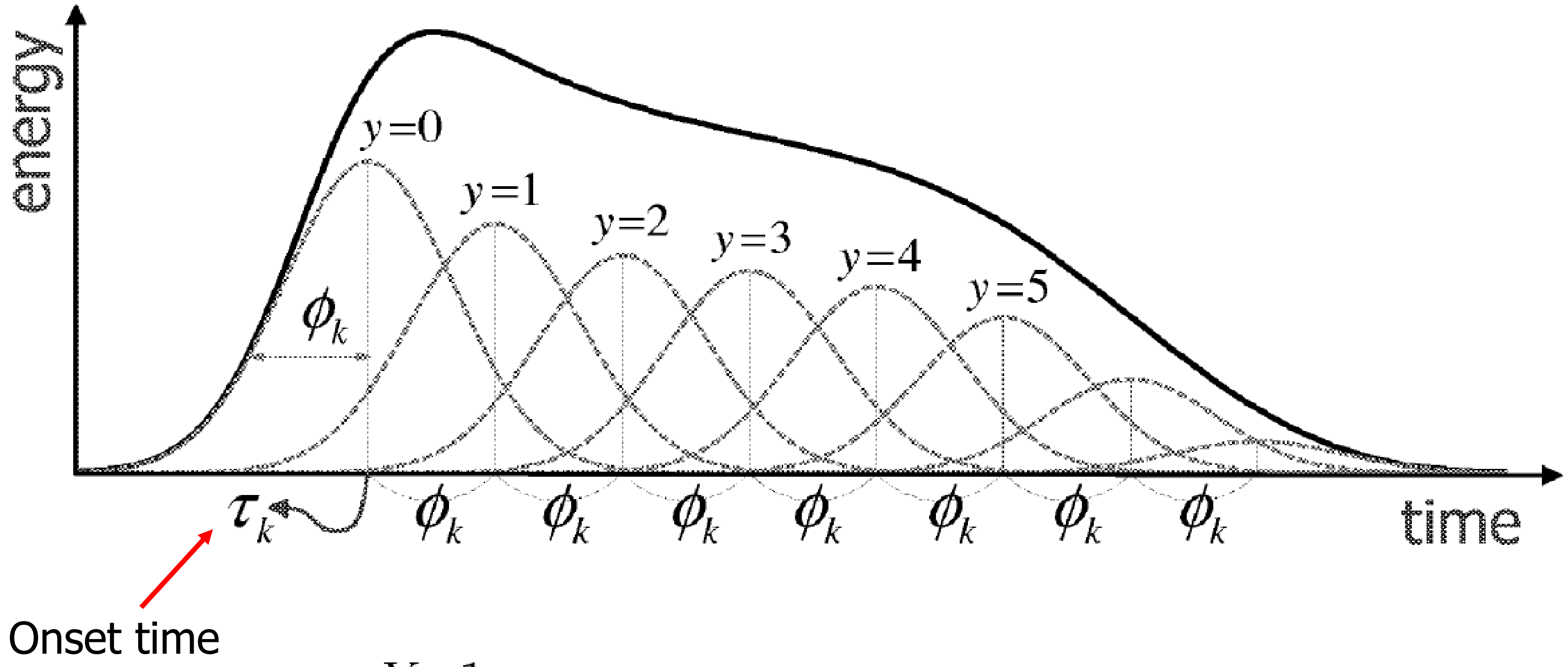
Pitch

The Model in A Single Frame



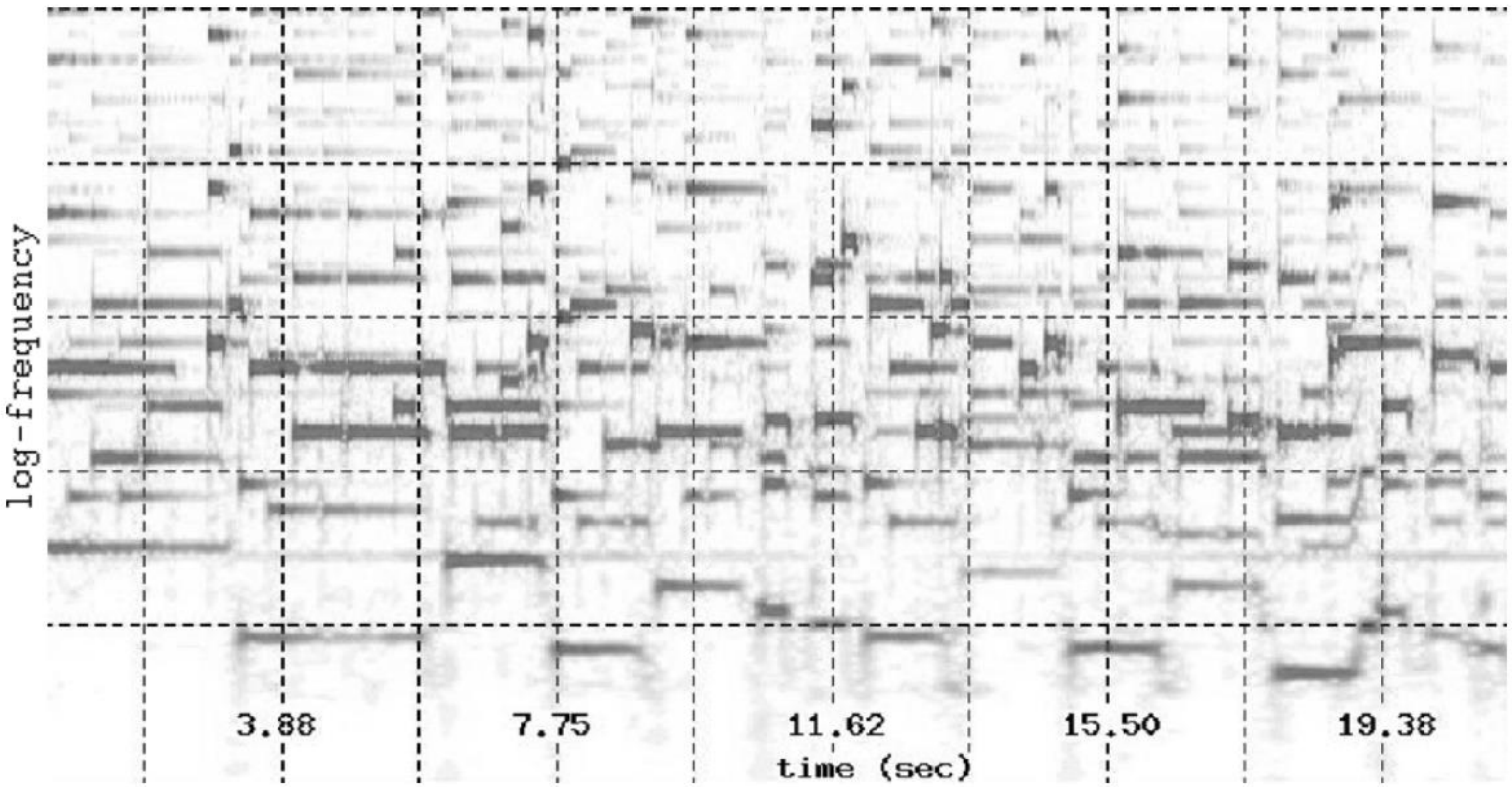
$$q_k(x,t;\mathbf{\Theta}) = w_k \sum_{n=1}^{N} \frac{v_{k,n} U_{k,n}(t)}{\sqrt{2\pi}\sigma_k} e^{-(x-\mu_k(t)-\log n)^2/2{\sigma_k}^2}$$

Harmonic Envelope



$$U_{k,n}(t) = \sum_{y=0}^{Y-1} \frac{u_{k,n,y}}{\sqrt{2\pi}\phi_{k,n}} \exp\left(-\frac{(t-\tau_k-y\phi_{k,n})^2}{2{\phi_{k,n}}^2}\right)$$

Reconstruction using HTC models



Activation weight of source k

$$\iint_D m_k(x,t)W(x,t)\log\frac{m_k(x,t)W(x,t)}{q_k(x,t;\Theta)}\mathrm{d}x\mathrm{d}t$$

The Unknowns

- Model parameters
  - Pitch, onset time, harmonic width, harmonic envelope over time, duration, etc.
- Latent variable
  - Activation weights of sources

- EM algorithm

Discussions

- Advantages
  - Very detailed model
  - Jointly estimates pitch, onset, duration, etc.
- Disadvantages
  - Model is very complicated

Classification-based Piano Note Transcription

[Poliner & Ellis, 2007]

- Train 88 (one-versus-all) SVM classifiers, one for each key of piano, from training audio frames
- Multi-label classification on each frame of the test audio

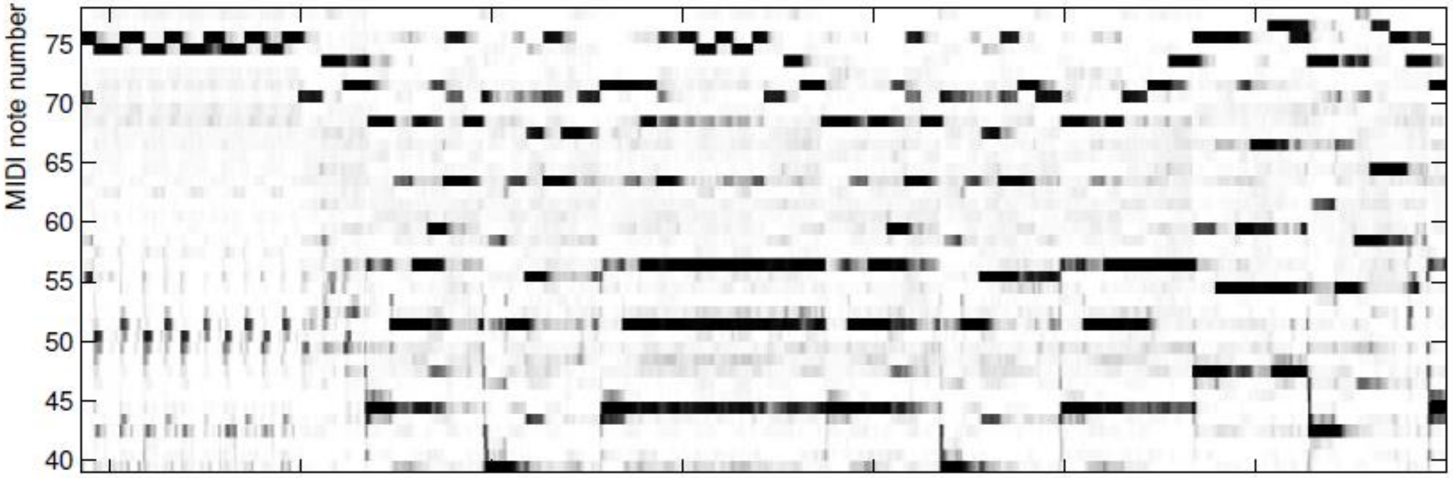
- Data: MIDI synthesized audio + Yamaha Disklavier playback grand piano
- Feature: a part of the magnitude spectrum

HMM Post Processing

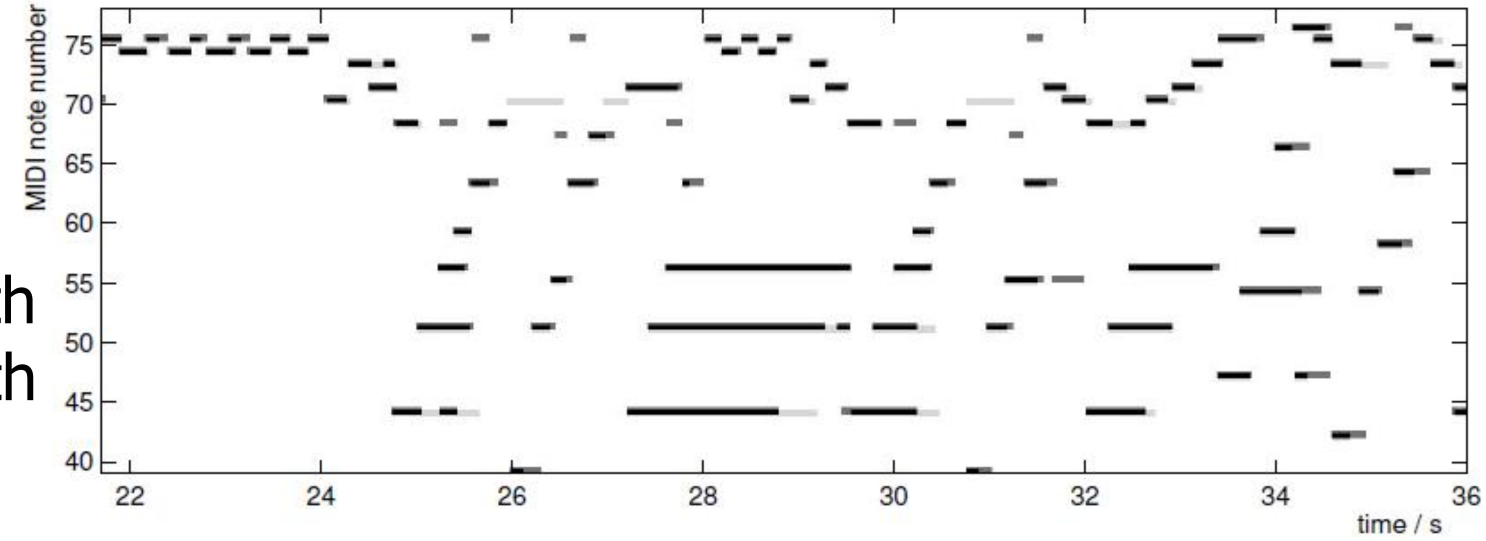
- 88 HMMs, one for each key
- 2 states: the pitch (key) is on/off
- Transition probability: learned from training data
- Observation probability (state likelihood): the probabilistic output of SVMs
- Viterbi algorithm to refine pitch estimates

HMM Post Processing Result

SVM probabilistic output, i.e. state likelihood



Refined pitch estimates, overlaid with ground-truth pitches

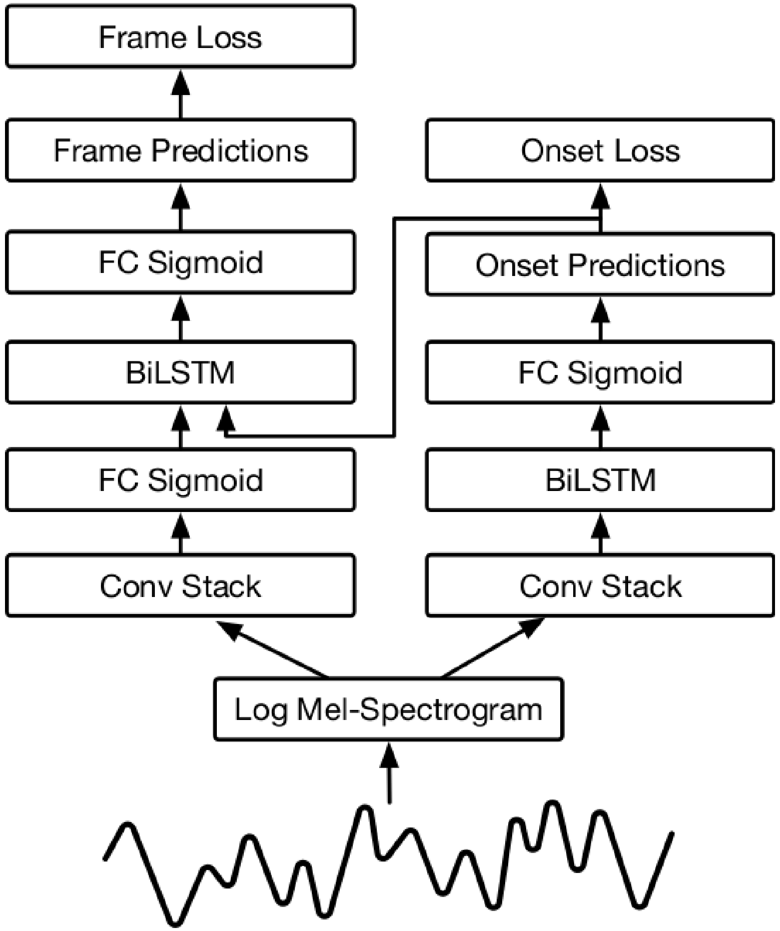


Discussions

- Advantages
  - The first classification-based transcription method
  - Simple idea
  - Easy to implement
- Disadvantages
  - The classification and post-processing of piano keys are performed totally independently
  - Induces more octave errors

Classification-based Piano Note Transcription

[Harthorne et al., 2018]



$$L_{frame} = \sum_{p=p_{min}}^{p_{max}} \sum_{t=0}^{T} CE\left(\mathbf{I}_{frame}(p,t), \mathbf{P}_{frame}(p,t)\right)$$

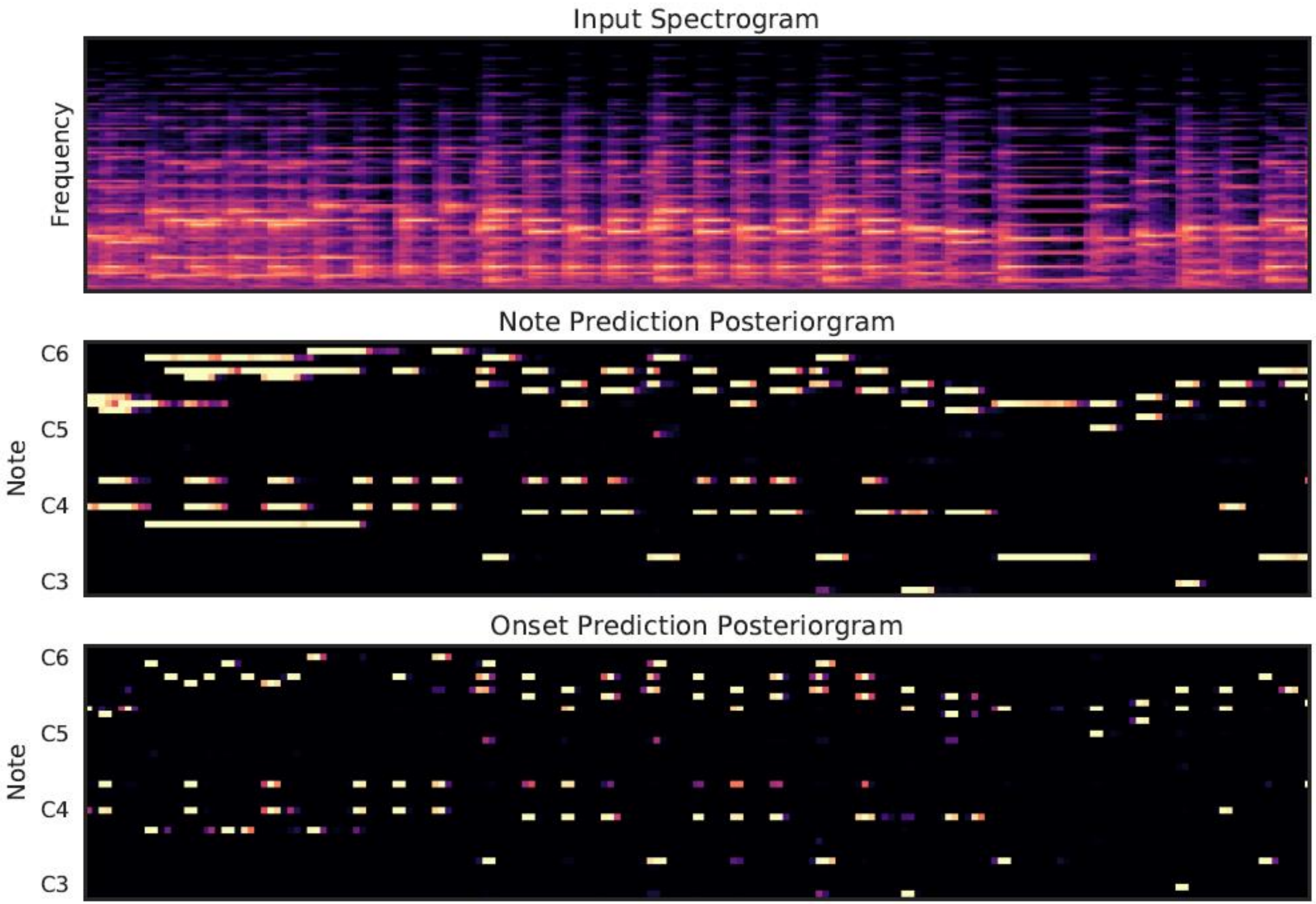
$$L_{onset} = \sum_{p=p_{min}}^{p_{max}} \sum_{t=0}^{T} CE\left(\mathbf{I}_{onset}(p,t), \mathbf{P}_{onset}(p,t)\right)$$

Model Training

- Training data
  - MAPS dataset: MIDI synthesized audio + Yamaha Disklavier playback grand piano

- Trained for 5 hours on 3 P100 GPUs

Example Output



Results

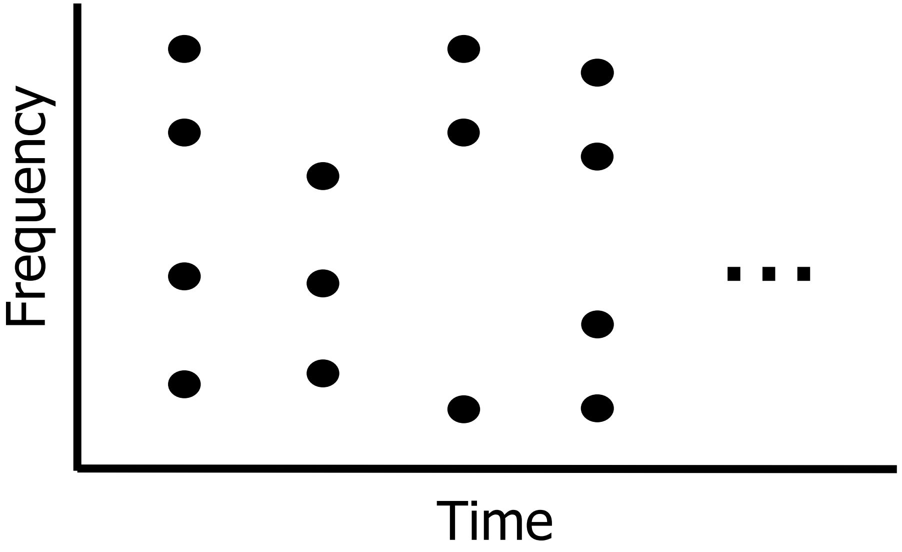
	Frame			Note			Note w/ offset			Note w/ offset & velocity		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Sigtia et al., 2016 [18]	71.99	**73.32**	72.22	44.97	49.55	46.58	17.64	19.71	18.38	—	—	—
Kelz et al., 2016 [13]	81.18	65.07	71.60	44.27	61.29	50.94	20.13	27.80	23.14	—	—	—
Melodyne (decay mode)	71.85	50.39	58.57	62.08	48.53	54.02	21.09	16.56	18.40	10.43	8.15	9.08
Onsets and Frames	**88.53**	70.89	**78.30**	**84.24**	**80.67**	**82.29**	**51.32**	**49.31**	**50.22**	**35.52**	**30.80**	**35.39**

Level 3: Multi-pitch Streaming

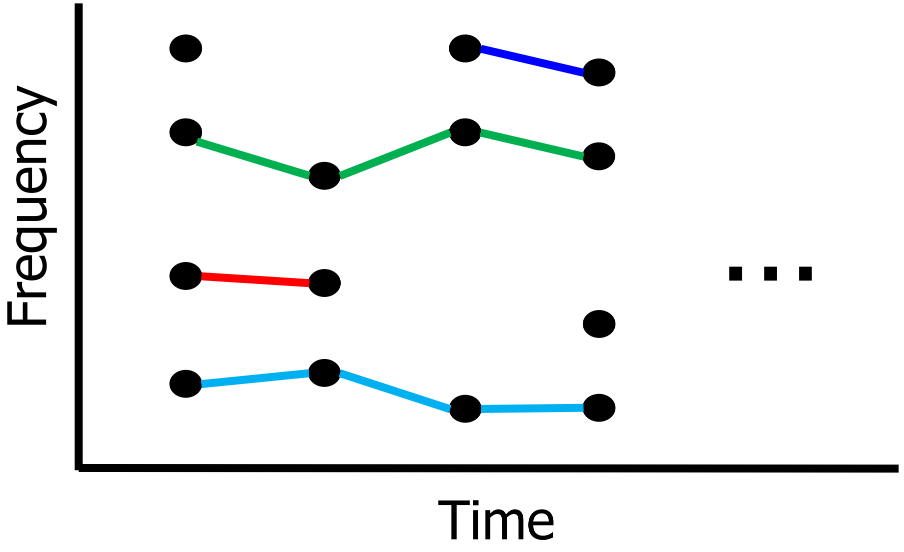
Estimate a pitch trajectory for each harmonic source

A 2-stage System

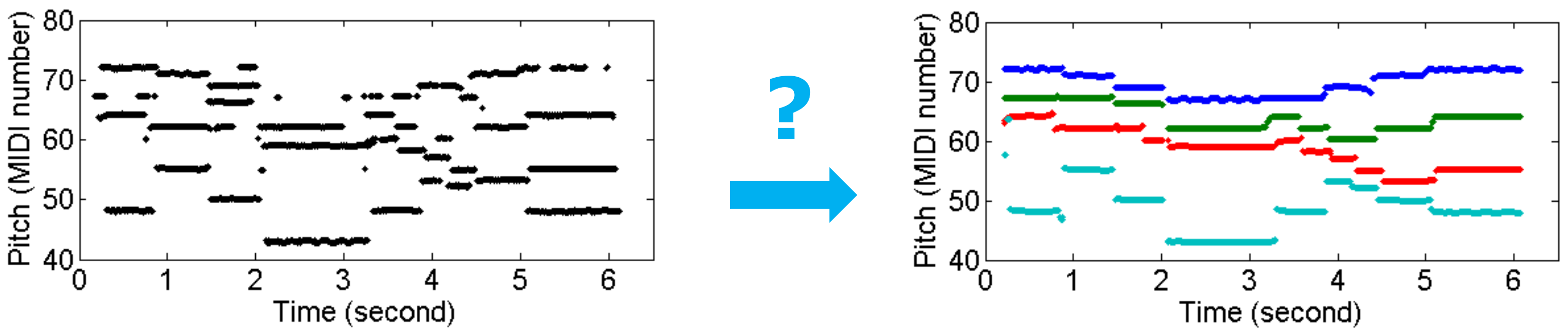
- Stage 1: Estimate pitches in each single time frame
  - [Duan et al., 2010]



- Stage 2: Connect pitch estimates across frames into pitch trajectories
  - [Duan et al., 2014]



How to Stream Pitches?



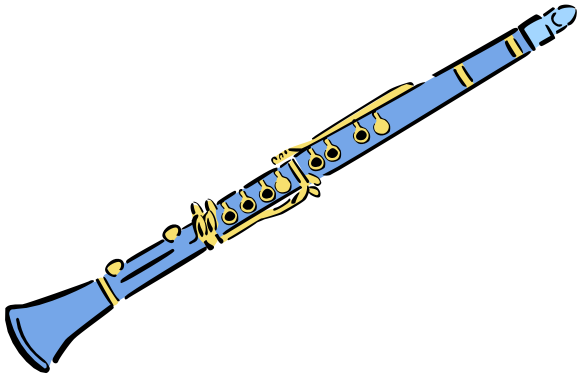
- Label pitches by pitch order in each frame, i.e. highest, second highest, third highest, …?
- Connect pitches by continuity?
  - Only achieves note tracking

Clustering Pitches by “Timbre”!

- Human use timbre to discriminate and track sound sources

“Timbre is that attribute of sensation in terms of which a listener can judge that two sounds having the same **loudness** and **pitch** are dissimilar.”

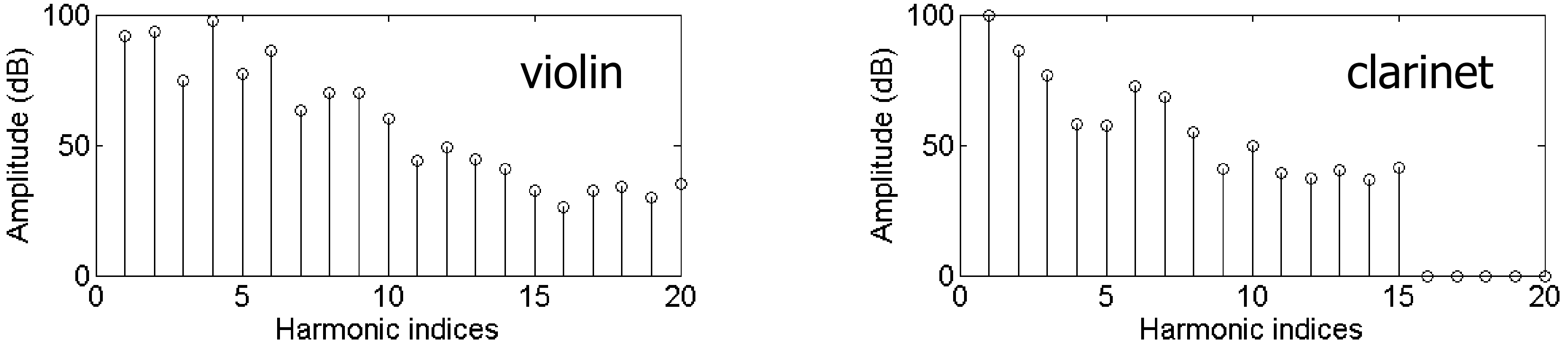
---- American Standards Association



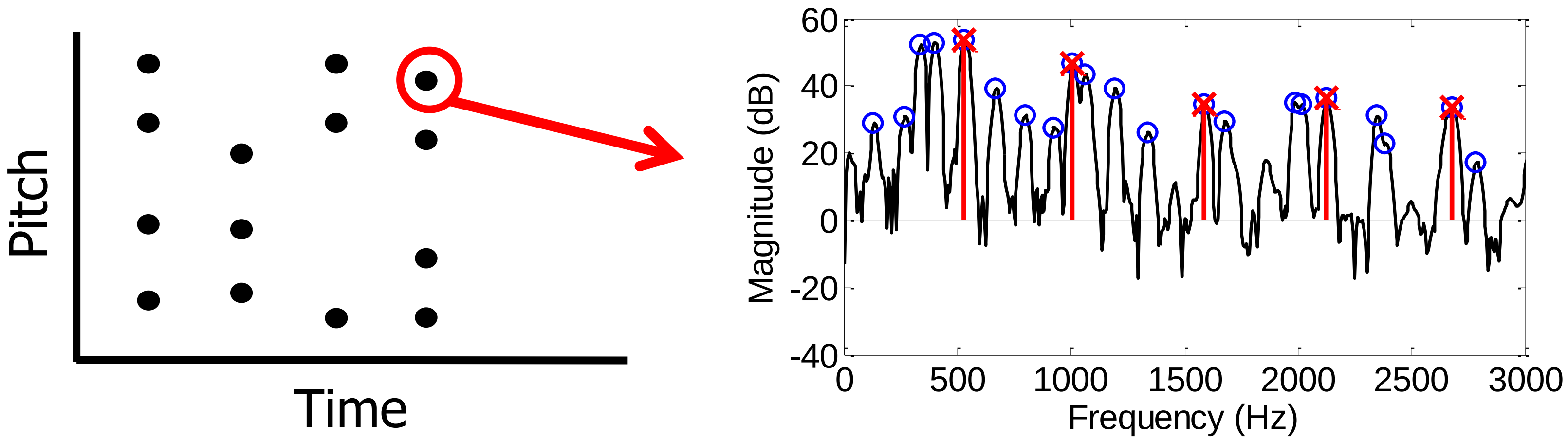
How to Represent Timbre?

- Harmonic structure

[Duan et. al. 2008]

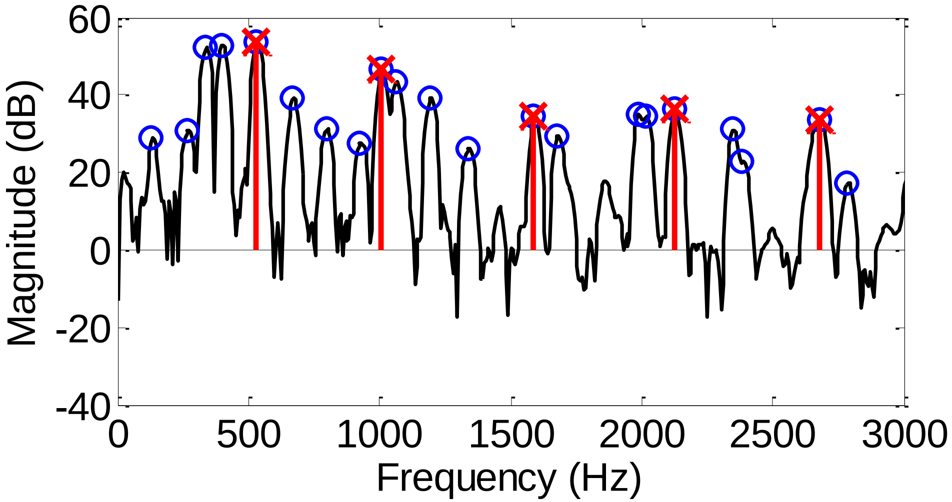


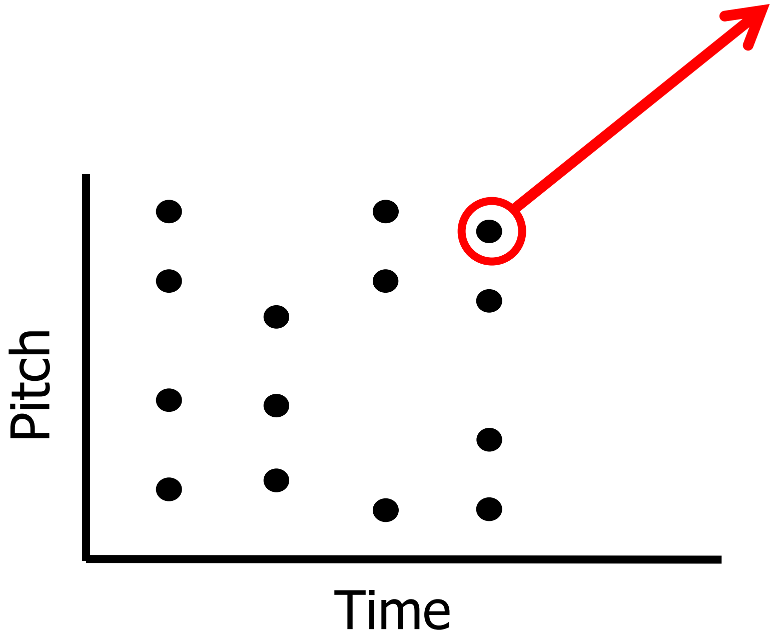
- Calculate for each pitch from the mixture



Timbre Feature for Talkers

- Characterizes talkers
- Calculated from mixture

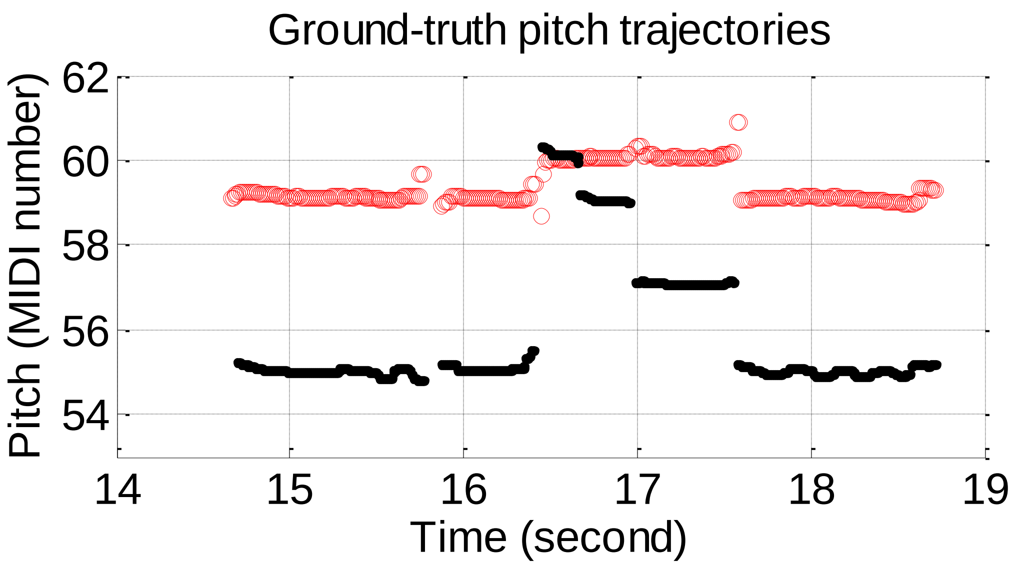


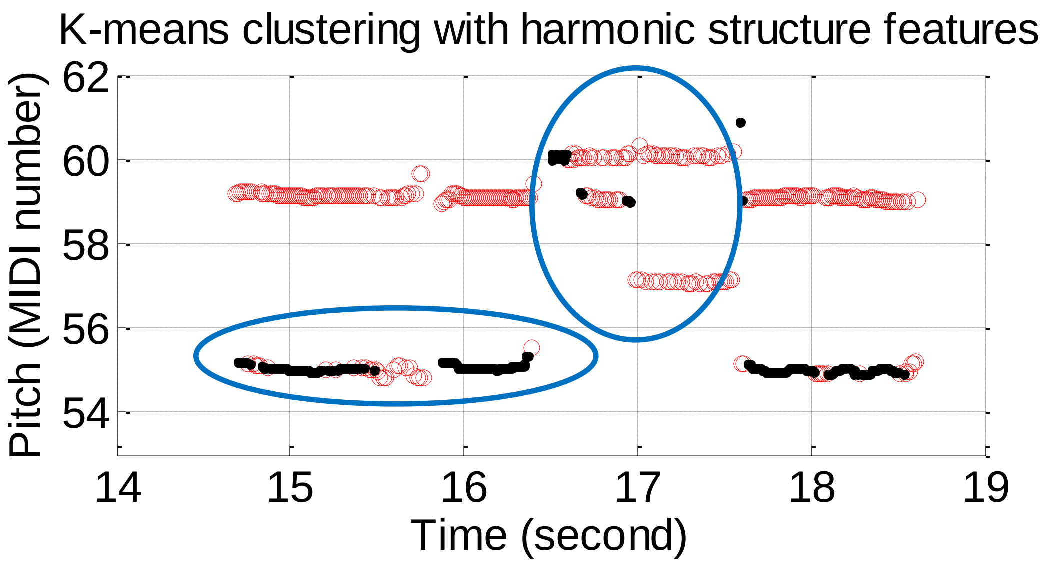


Discrete Cosine Transform

Uniform Discrete Cepstrum (UDC)

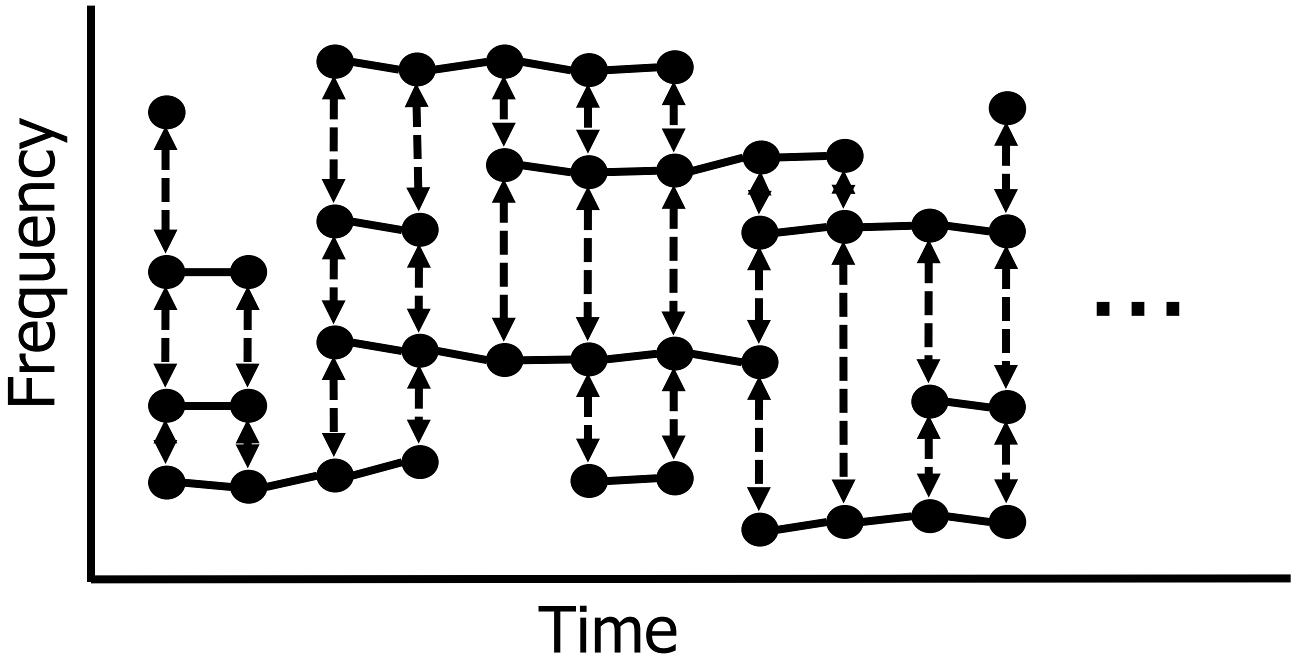
Clustering by timbre is not enough





Use Pitch Locality Constraints

- Cannot-link: between simultaneous pitches (only for monophonic instruments)
- Must-link: between pitch estimates close in both time and frequency



Constrained Clustering

- Objective: minimize timbre inconsistency
- Constraints: pitch locality
  - Inconsistent constraints: caused by incorrect pitch estimates, interweaving pitch trajectories, etc.
  - Heavily constrained: nearly every pitch estimate is involved in at least one constraint

- Algorithm: iteratively update the clustering s.t.
  - The objective monotonically decreases
  - The set of satisfied constraints monotonically expands

The Proposed Algorithm

- f: objective function; C: all constraints;
- Π_n: clustering in n-th iteration;
- C_n: {constraints satisfied by Π_n} ;

1. $n \leftarrow 0$; Start from an initial clustering $<\Pi_0, C_0>$;
2. $n \leftarrow n + 1$; Find a new clustering Π_n such that $f(\Pi_{n-1}) > f(\Pi_n)$, and Π_n also satisfies C_{n-1};
3. C_n = {constraints satisfied by Π_n}; so $C_{n-1} \subseteq C_n$

- It converges to some local minimum $< \Pi', C' >$.

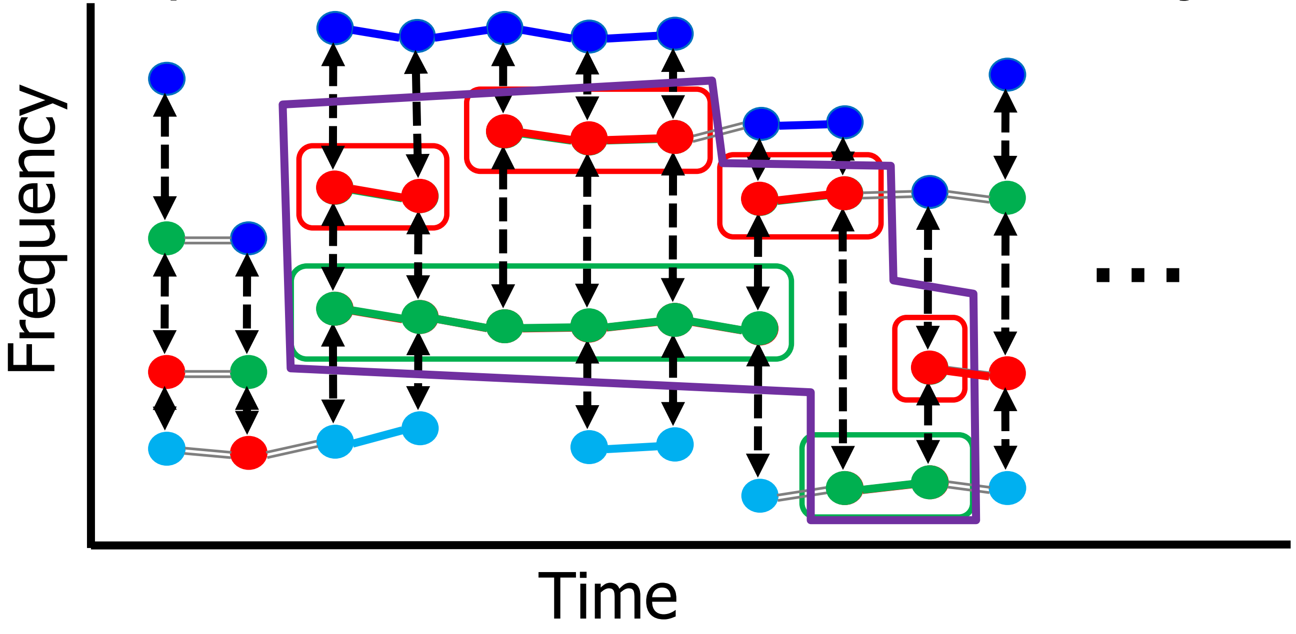
$$f(\Pi_0) > f(\Pi_1) > \cdots > f(\Pi')$$

$$C_0 \subseteq C_1 \subseteq \cdots \subseteq C'$$

Find A New Clustering to…

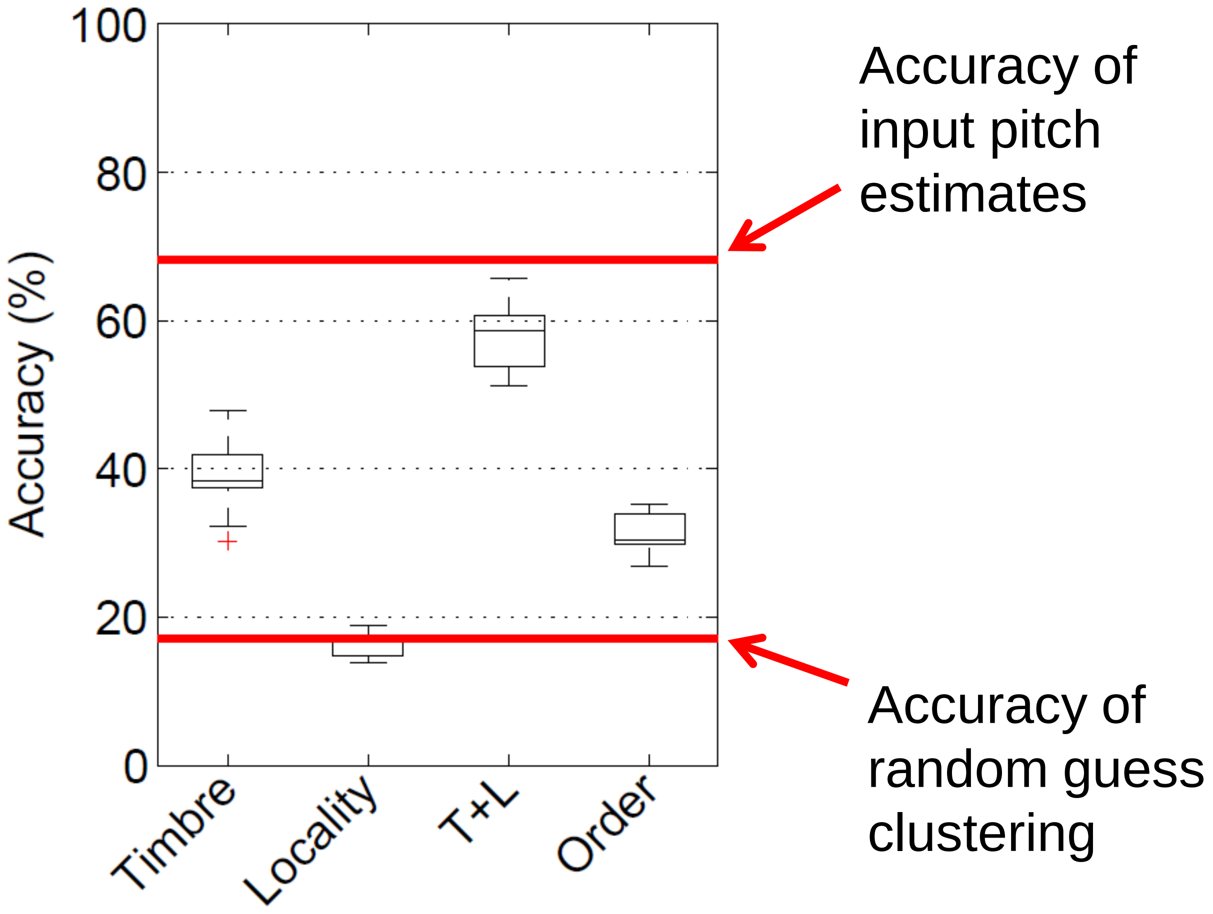
1. Decrease the objective function
2. Satisfy satisfied constraints

- Swap set: a connected graph between two clusters by already satisfied constraints
- One more must link is satisfied now
- Try all swap sets to find one that decreases objective



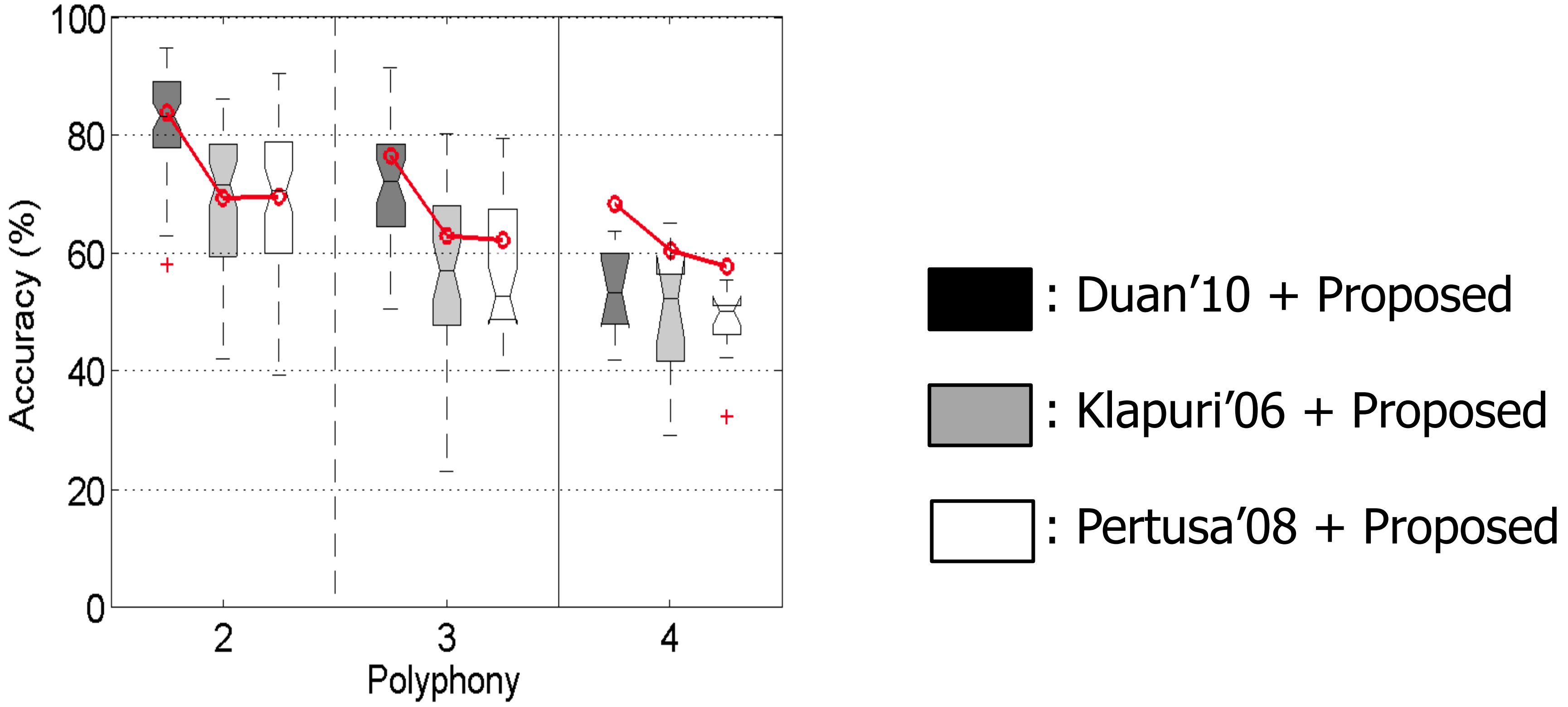
Timbre Objective & Locality Constraints

- Results on 10 quartets played by violin, clarinet, saxophone and bassoon



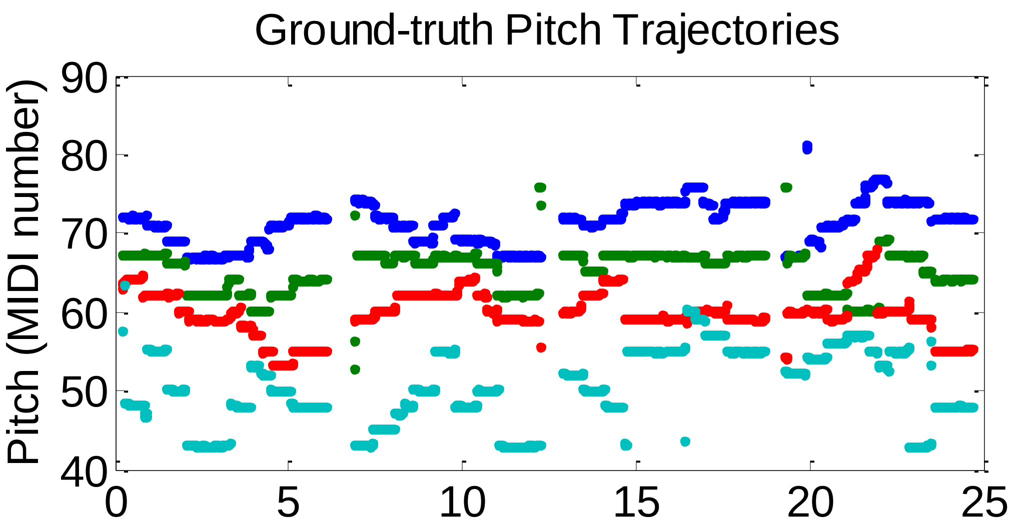
Works with Different MPE Methods

- Results on 60 duets, 40 trios, and 10 quartets



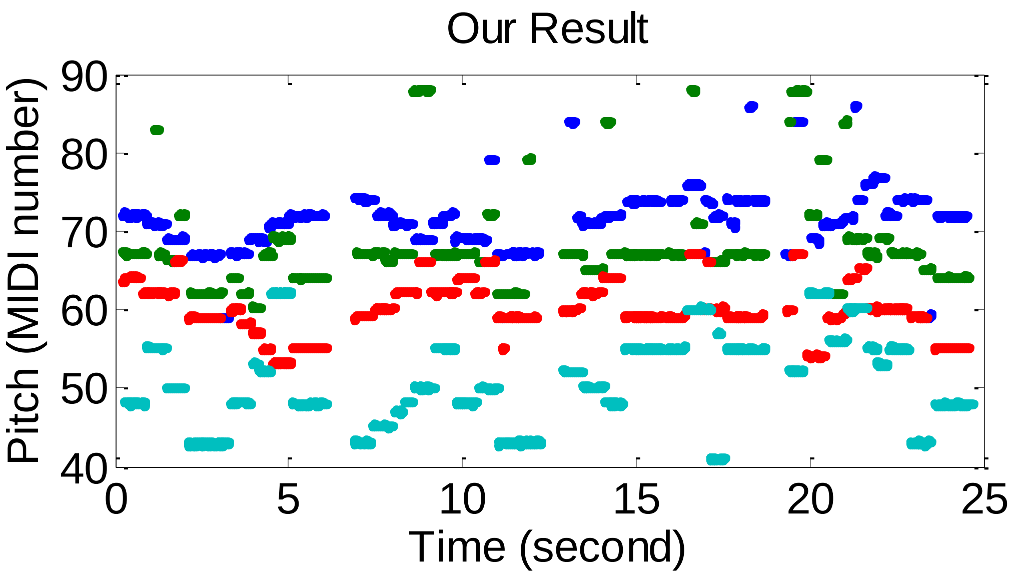
Example on Music

Original violin (blue)



Original clarinet (green)

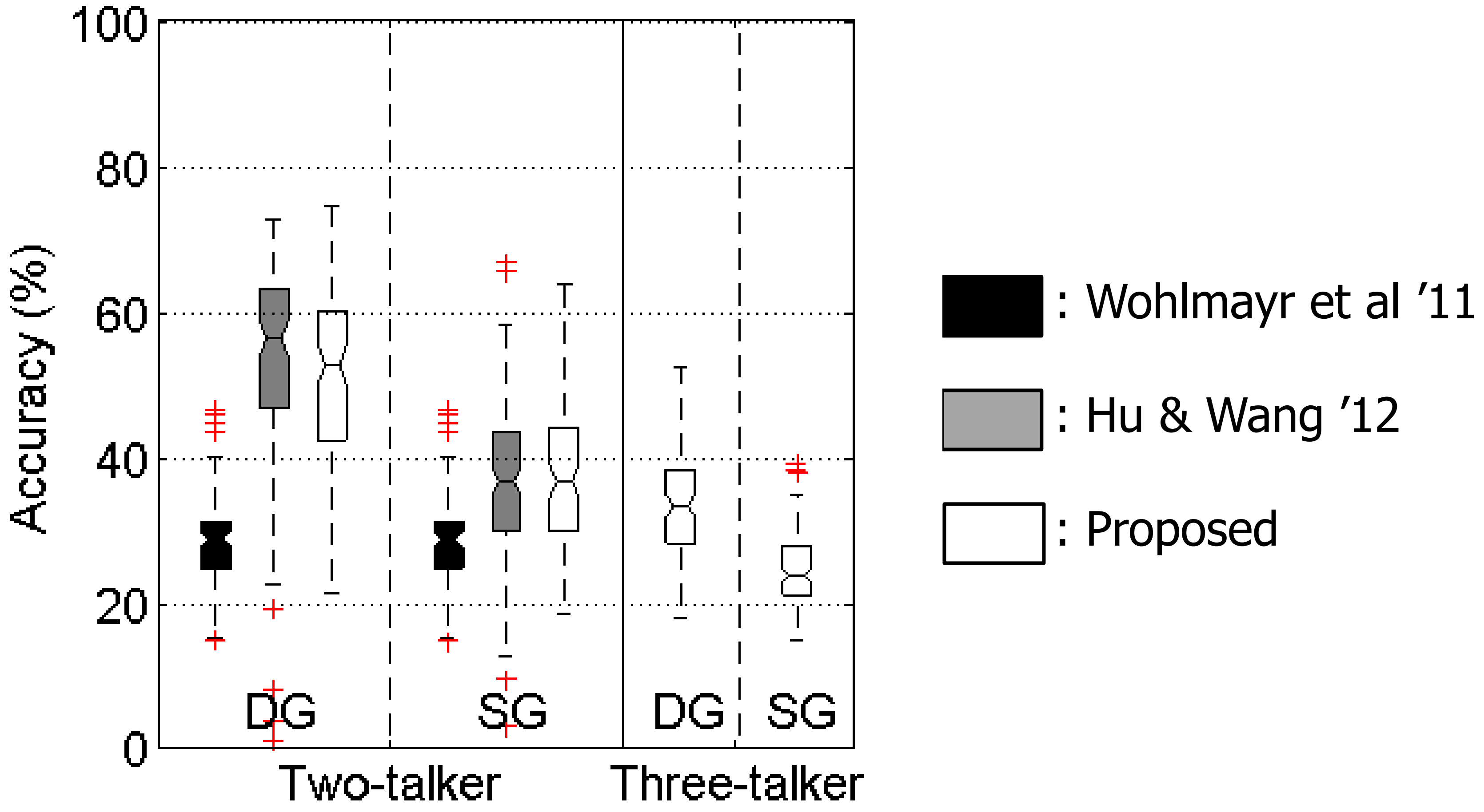
Separated violin (blue)



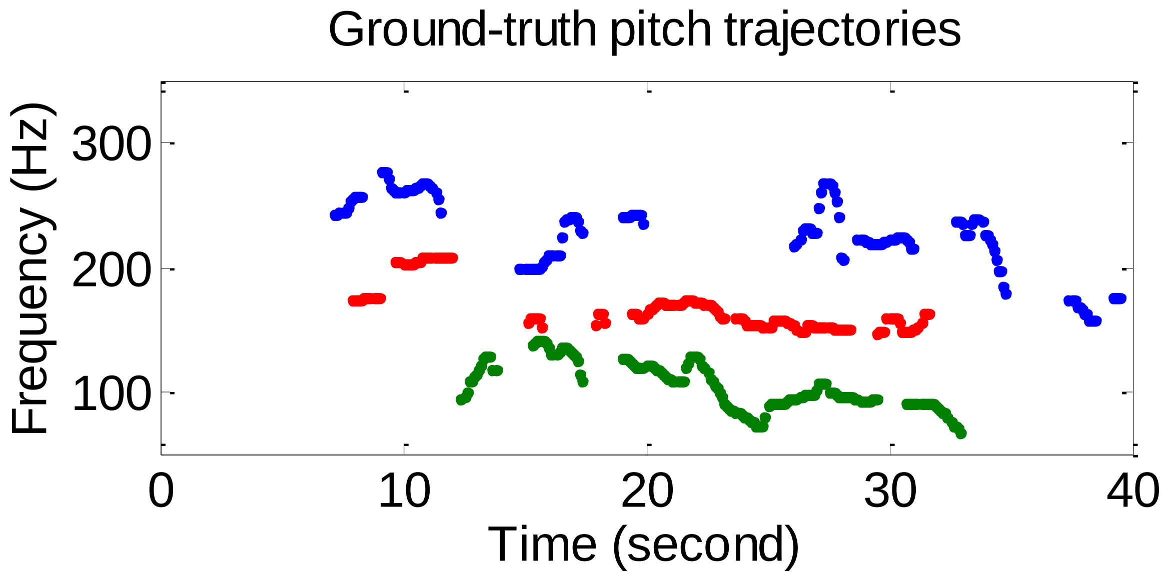
Separated clarinet (green)

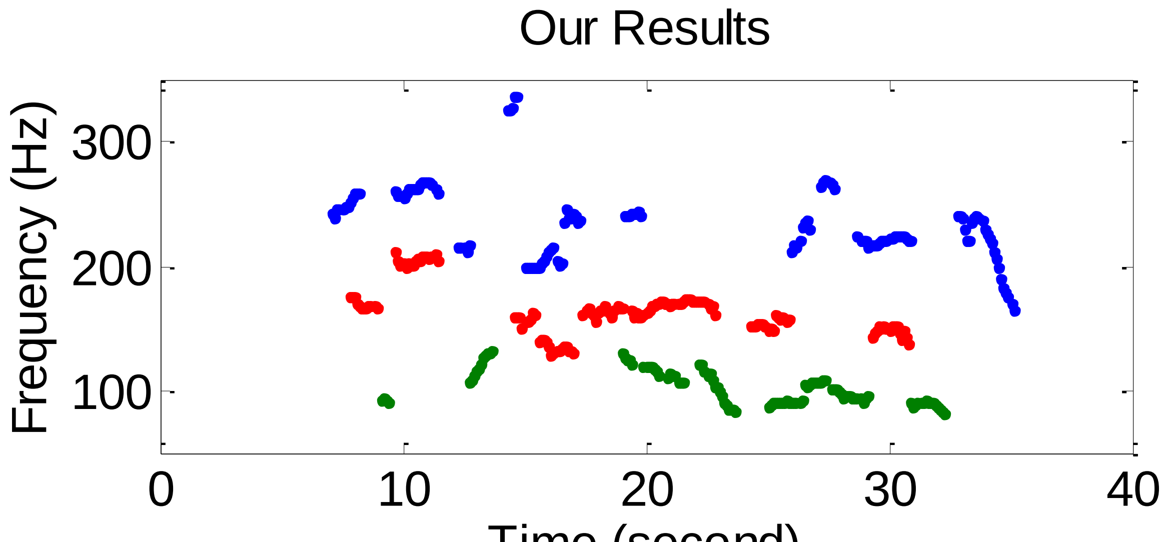
Comparisons on Speech

- 400 2-talker and 3-talker speech excerpts



Example on Speech





Discussions

- Advantages:
  - Able to stream pitches across notes
  - Considers both timbre and pitch location info

- Disadvantages:
  - Algorithm is slow and complicated.
  - Constraints are binary.
  - Cannot deal with polyphonic instruments e.g. piano and guitar.