Music Pitch Analysis

Zhiyao Duan

Associate Professor of ECE and CS

University of Rochester

Presentation at WiSSAP 2023, IIT Kanpur, December 18-21, 2023

Outline

- Basic Concepts of Pitch
- Single Pitch Detection
- Multi-Pitch Analysis

Pitch (ANSI 1994 Definition)

- That attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high. Pitch **depends mainly on the frequency** content of the sound stimulus, but **also depends on the sound pressure and waveform** of the stimulus

- (Operational) A sound has a certain pitch if it can be reliably matched to a sine tone of a given frequency at 40 dB SPL

Pitch and Intensity

- Stevens Rule
  - The pitch of low frequency (below 1000Hz) sine tones decreases with increasing intensity -- (low loud sounds go flat)
  - The pitch of high frequency tones (over 3000 Hz) increases with intensity -- (high loud sounds go sharp)

220Hz

7040Hz

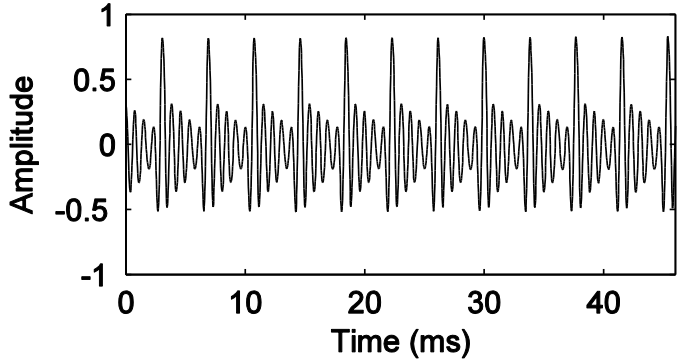
Harmonic Sound

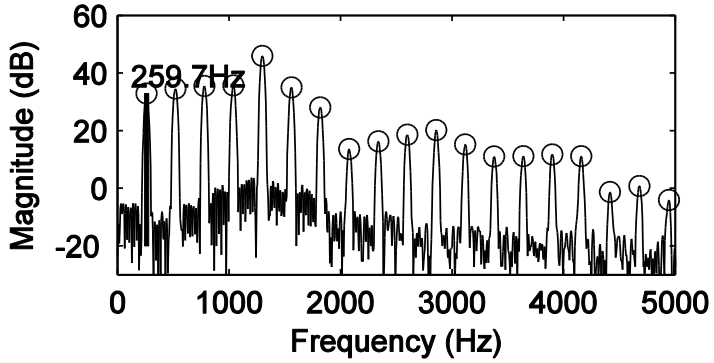
- A sound with strong sinusoidal components at integer multiples of a fundamental frequency. These components are called **harmonics** or **overtones**.

- Harmonic sounds are the sounds that may give a perception of “pitch”.

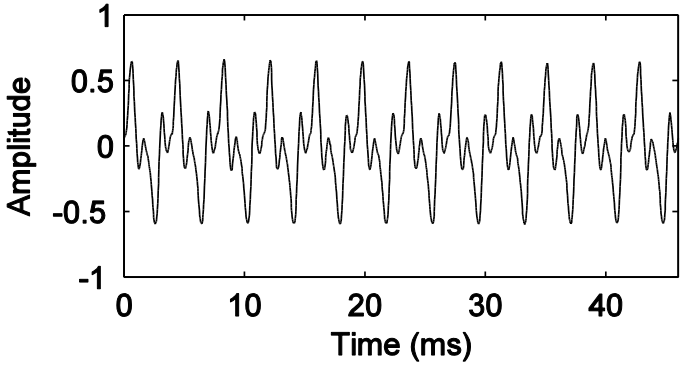
Classify Sounds by Harmonicity

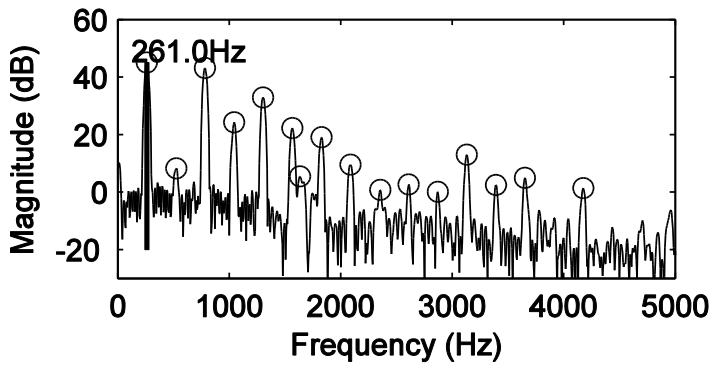
- Sine wave
- Strongly harmonic





Oboe

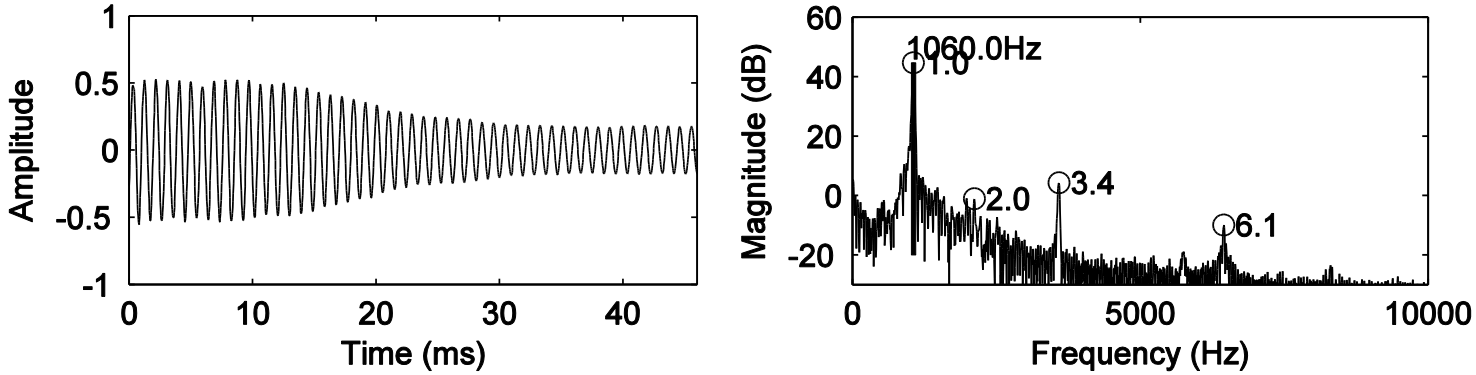




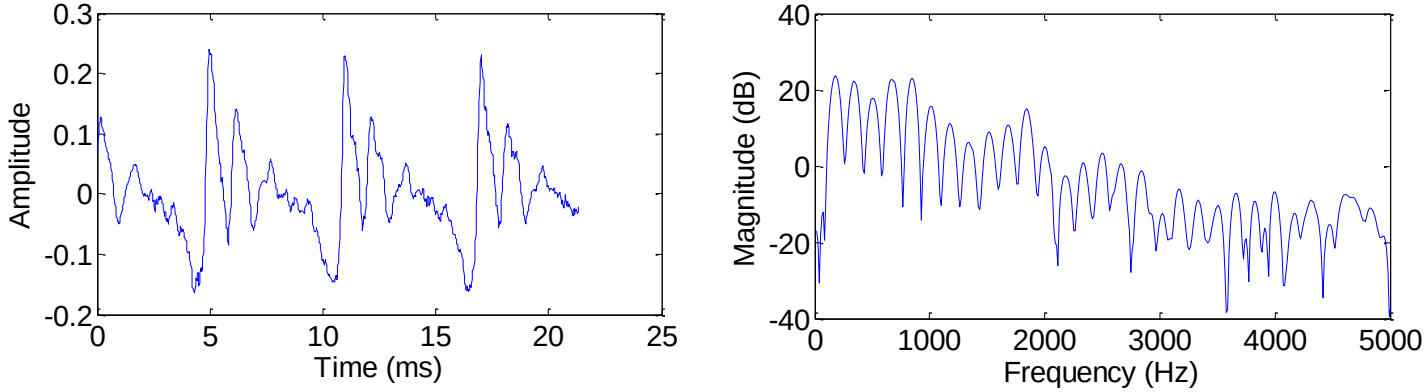
Clarinet

Classify Sounds by Harmonicity

- Somewhat harmonic (quasi-harmonic)



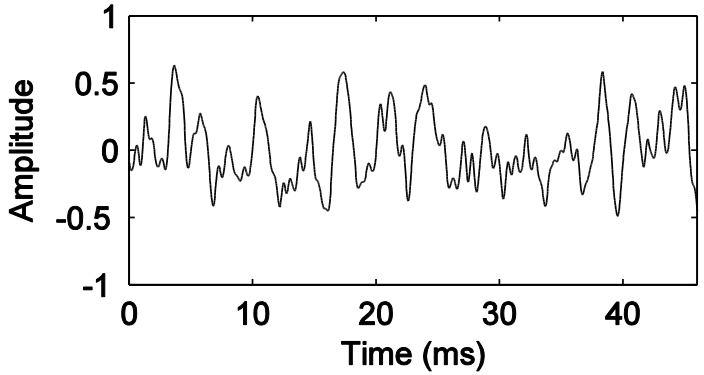
Marimba

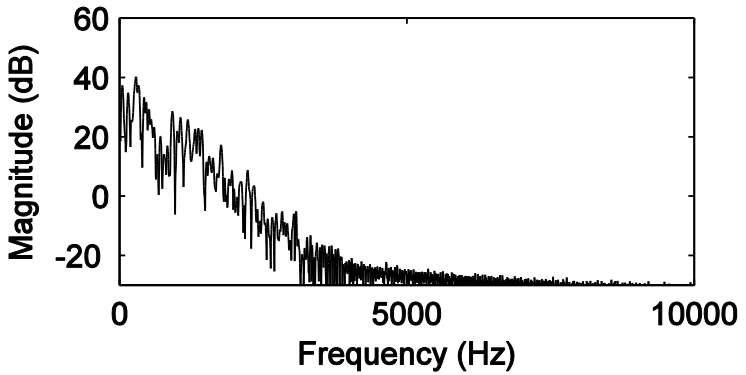


Human voice

Classify Sounds by Harmonicity

- Inharmonic





Gong

Sounds	Instrument family	Instruments
Harmonic	Woodwind	Piccolo, flute, oboe, clarinet, bassoon, saxophone
	Brass	Trumpet, horn, euphonium, trombone, tuba
	Arco string	Violin, viola, cello, double bass
	Pluck string	Piano, guitar, harp, celesta
	Vocal	Voiced phonemes
Quasi-harmonic	Pitched percussive	Timpani, marimba, vibraphone, xylophone
Inharmonic	Non-pitched percussive	Drums, cymbal, gong, tambourine

(from Anssi Klapuri, and Manuel Davy, editors. *Signal Processing Methods for Music Transcription*. Springer, 2006.)

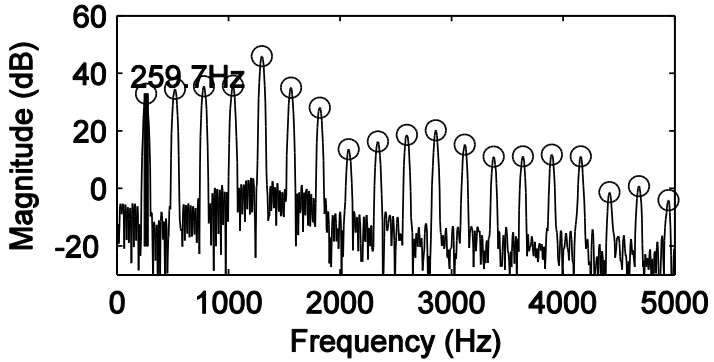
What determines pitch?

- Complex tones
  - Strongest frequency?
  - Lowest frequency?
  - Something else?

- Let’s listen and explore…

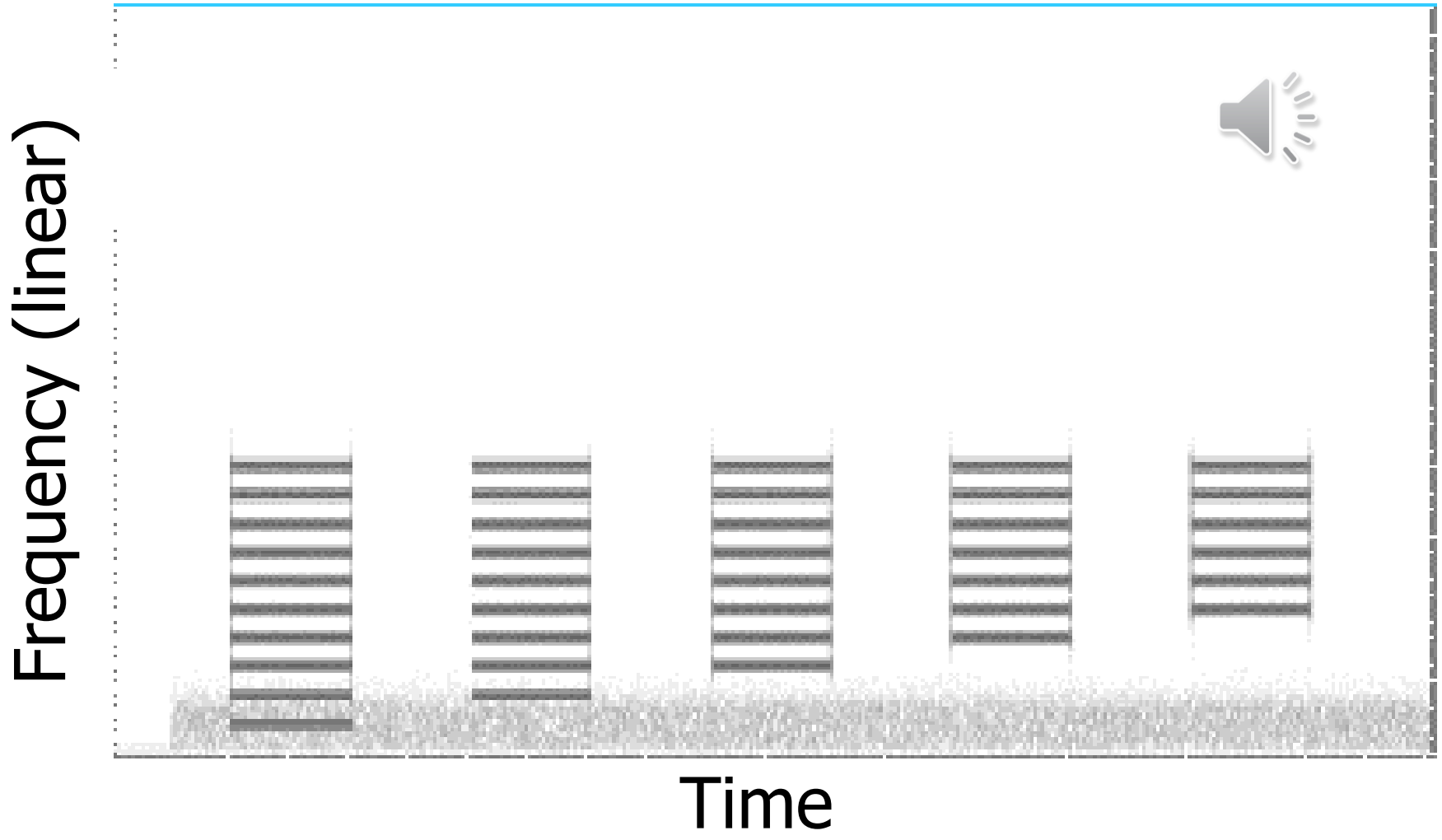
Hypothesis

- Pitch is determined by the lowest strong frequency component in a complex tone



Oboe

The Missing Fundamental



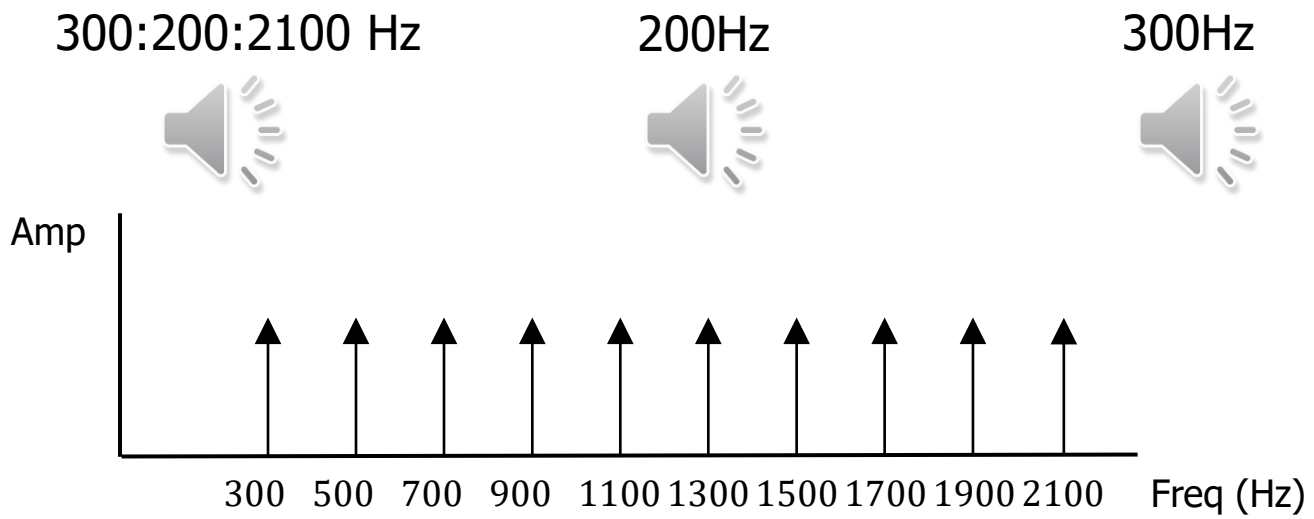
Hypothesis

- Pitch is determined by the lowest strong frequency component in a complex tone

- The case of the missing fundamental proves that it’s not always so

Hypothesis – “It’s complicated”

- by the loudest frequency
- by the common frequency that divides other frequencies
- by the space between regularly spaced frequencies



Pitch vs. F0

- A perceptual attribute, so subjective
- Only defined for (quasi) harmonic sounds
  - Harmonic sounds are periodic, and the period is 1/F0.
- Can be reliably matched to fundamental frequency (F0)
  - In computer audition, people do not often discriminate pitch from F0
- F0 is a physical attribute, so objective

Pitch and Music

- How do we tune pitch in music?
- How do we represent pitch in music?
- How do we represent the relation of pitches in music?

Equal Temperament

- Octave is a relationship by the power of 2
- There are 12 half-steps in an octave

number of half-steps
from the reference pitch

$$f(n) = 2^{\frac{n}{12}} f_{\text{ref}}$$

frequency of
desired pitch

frequency of the
reference pitch

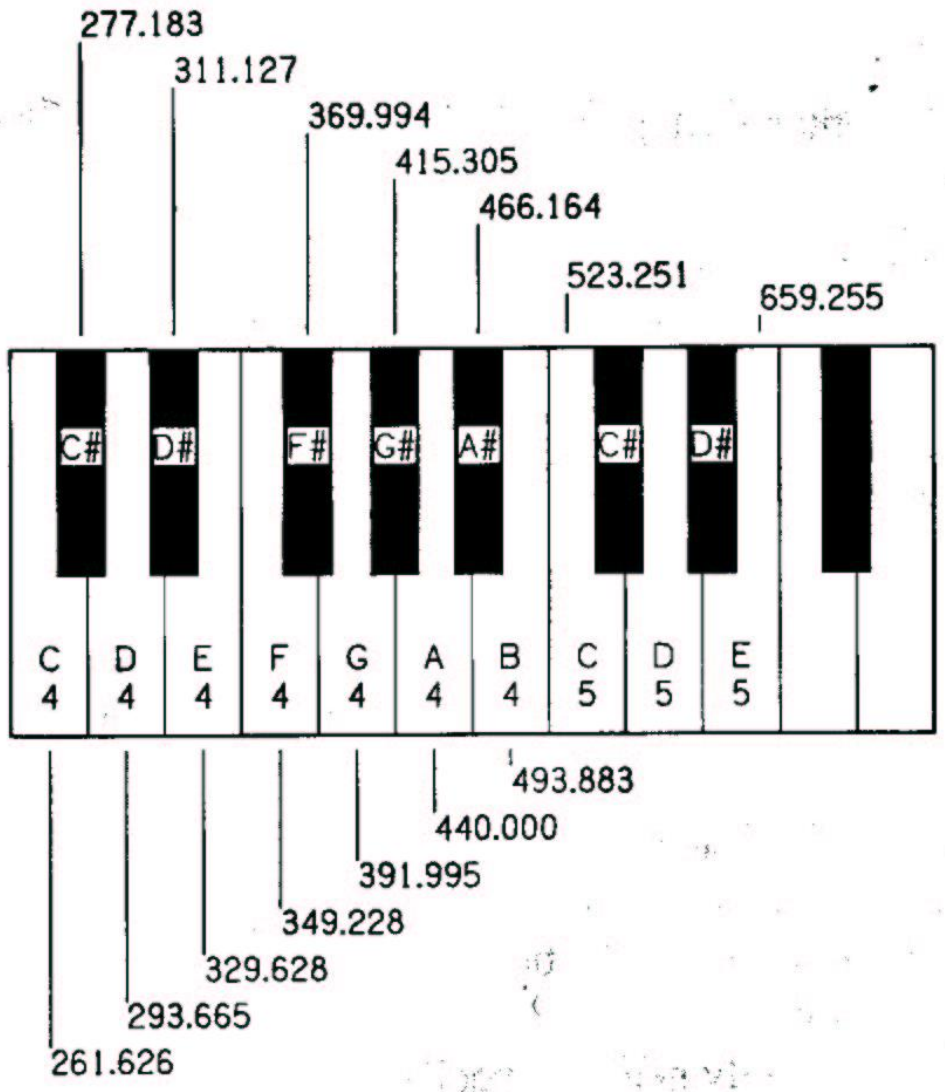
Measurement

- 100 Cents in a half step
- 2 half steps in a whole step
- 12 half steps in an octave

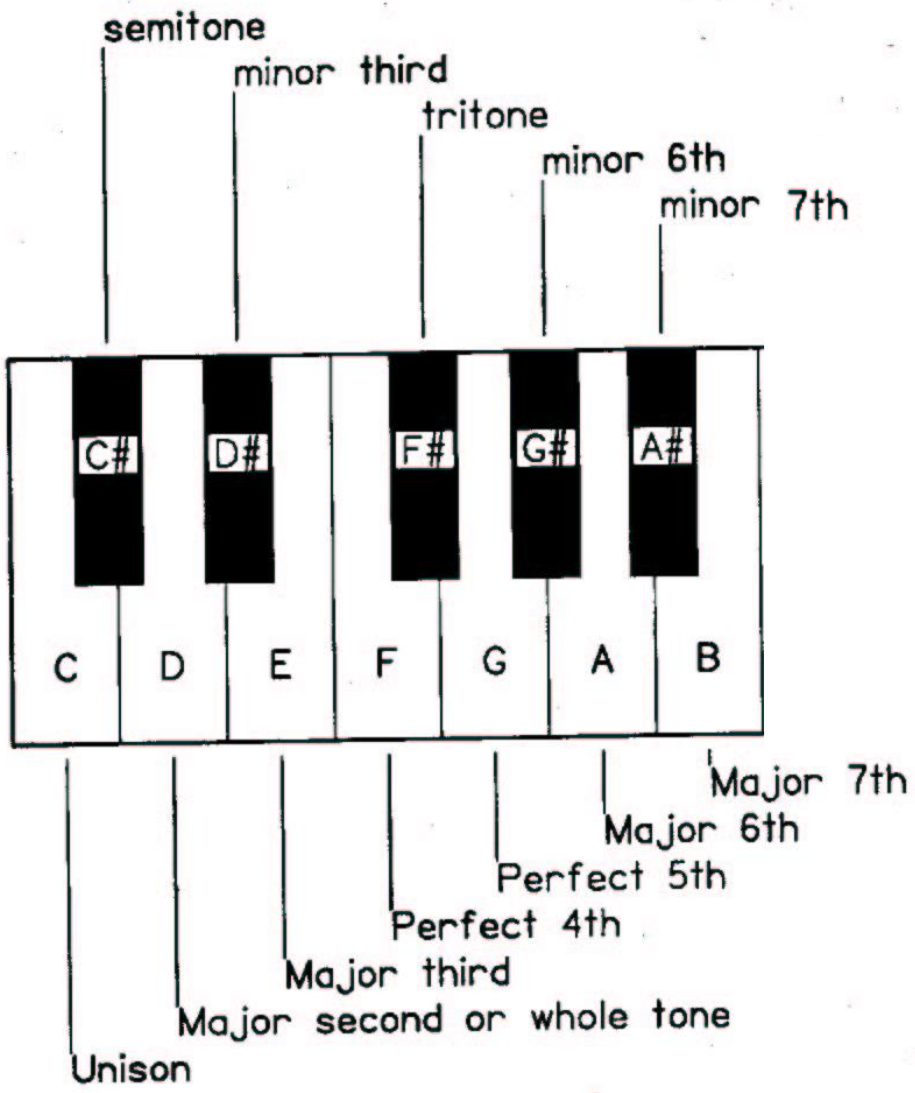
Number of cents

$$c = 1200 \log_2 \left(\frac{f}{f_{\text{ref}}} \right)$$

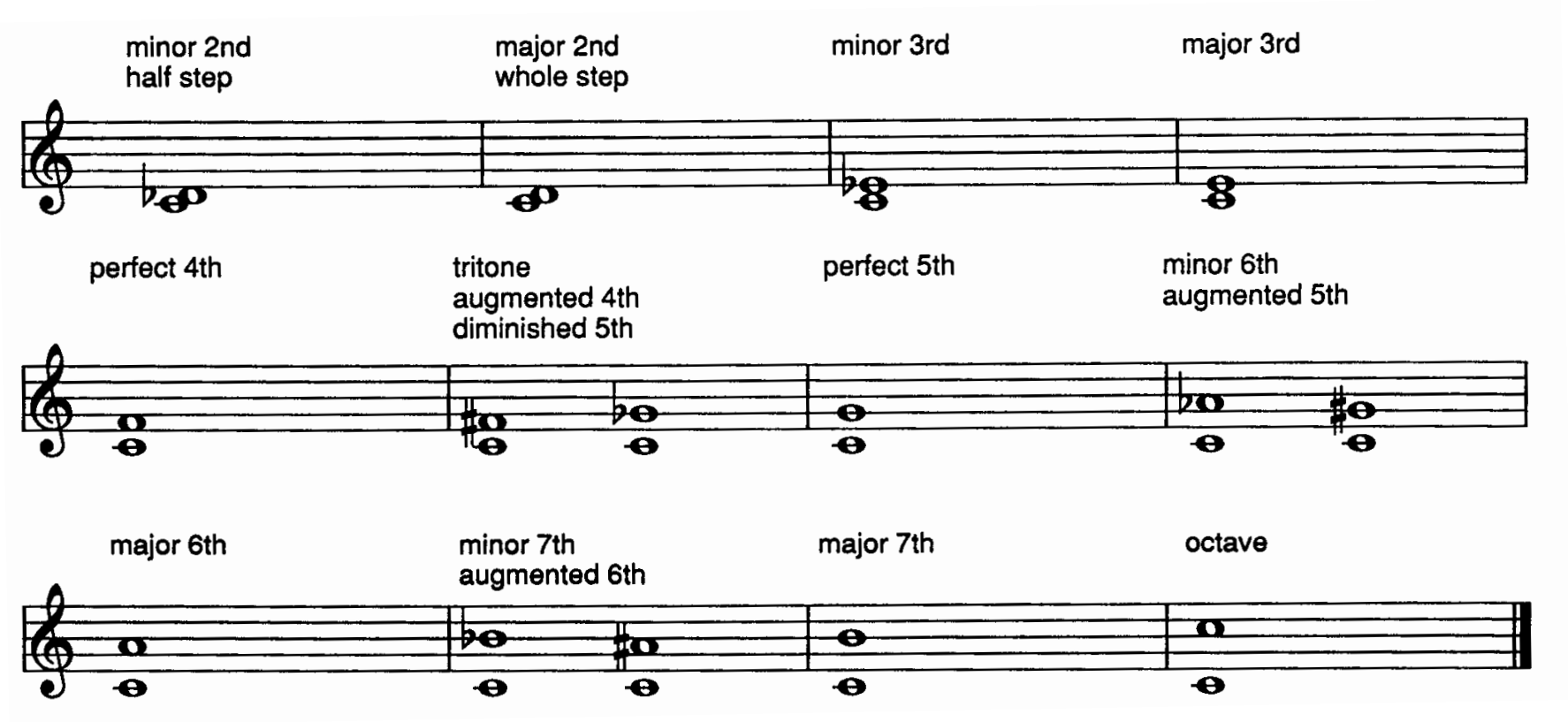
A=440 Equal Temperament Tuning



Musical Intervals (from C)

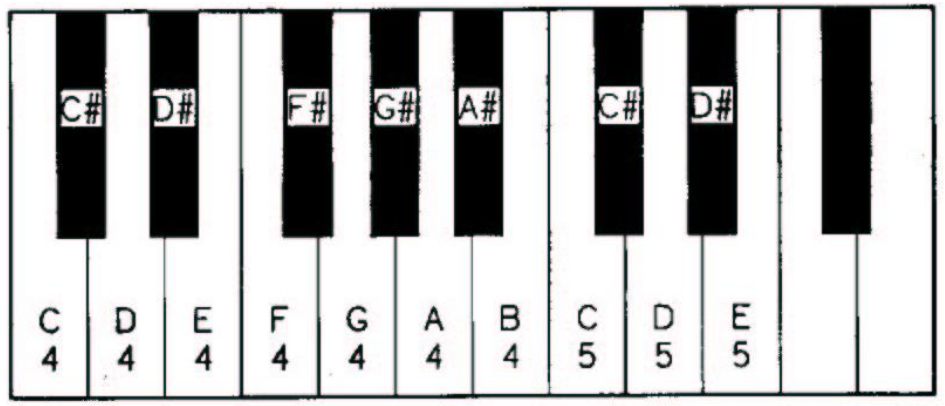


Interval Names



Some Magic

Half-steps: 0 1 2 3 4 5 6 7 8 9 10 11 12



C → C: 12 half-steps, $2^{\frac{12}{12}} = \frac{2}{1}$

C → G: 7 half-steps, $2^{\frac{7}{12}} = 1.4983 \approx \frac{3}{2}$

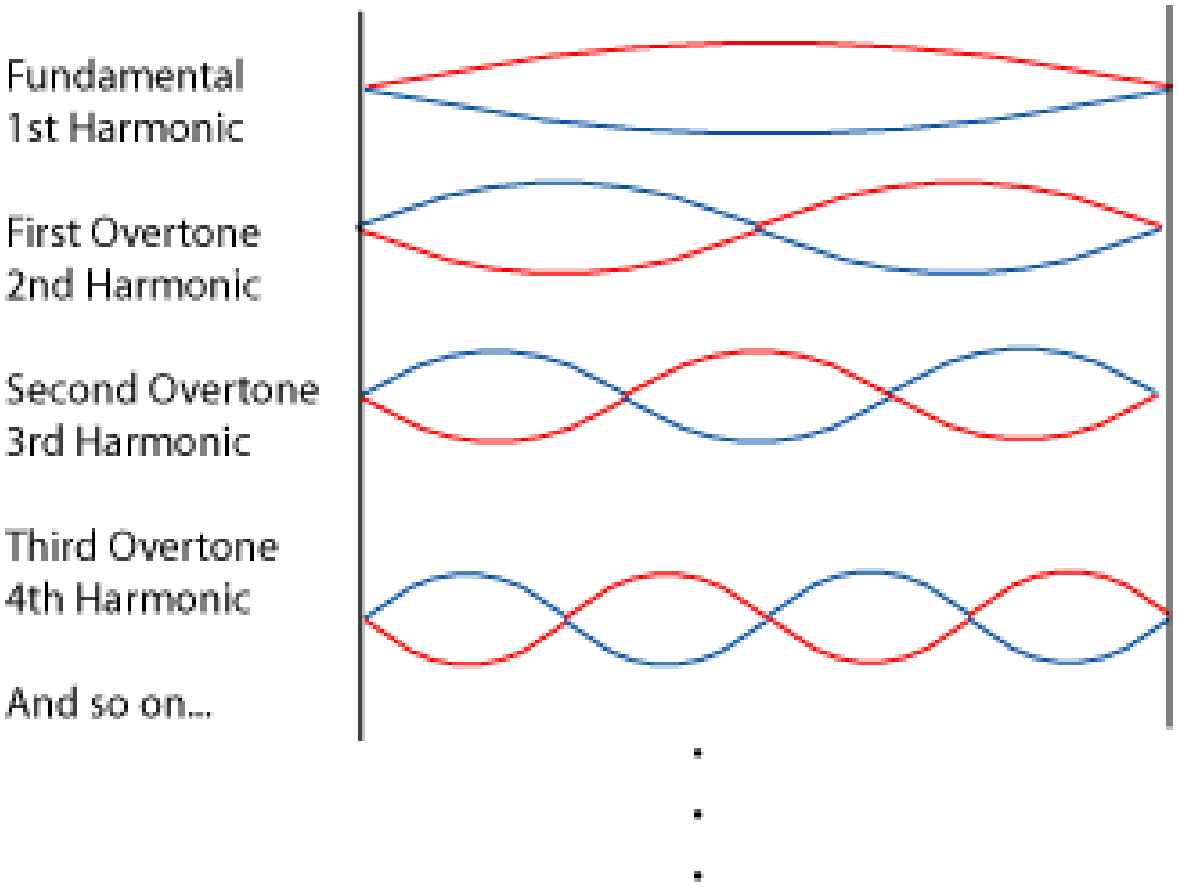
C → F: 5 half-steps, $2^{\frac{5}{12}} = 1.3348 \approx \frac{4}{3}$

C → E: 4 half-steps, $2^{\frac{4}{12}} = 1.2599 \approx \frac{5}{4}$

Are these just coincidence?

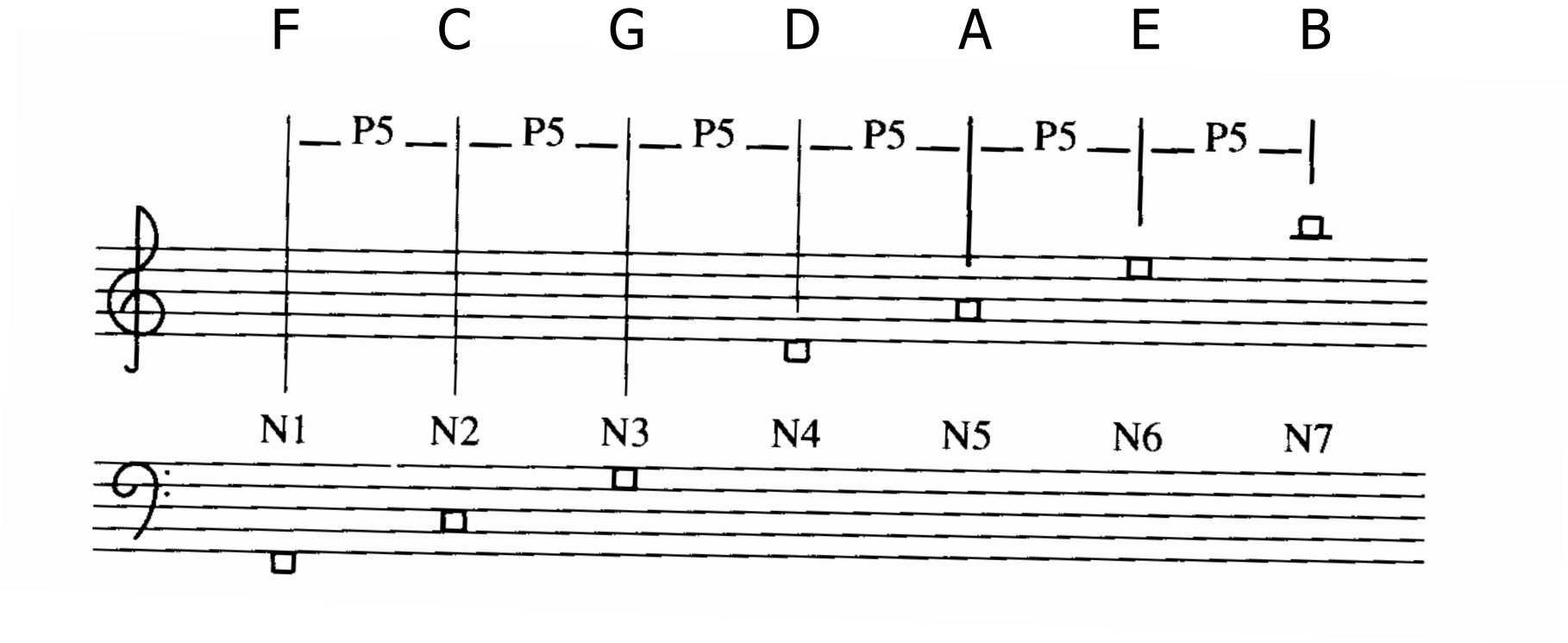
Related to Standing Waves

- How about defining pitches this way, so that they sound more harmonic?

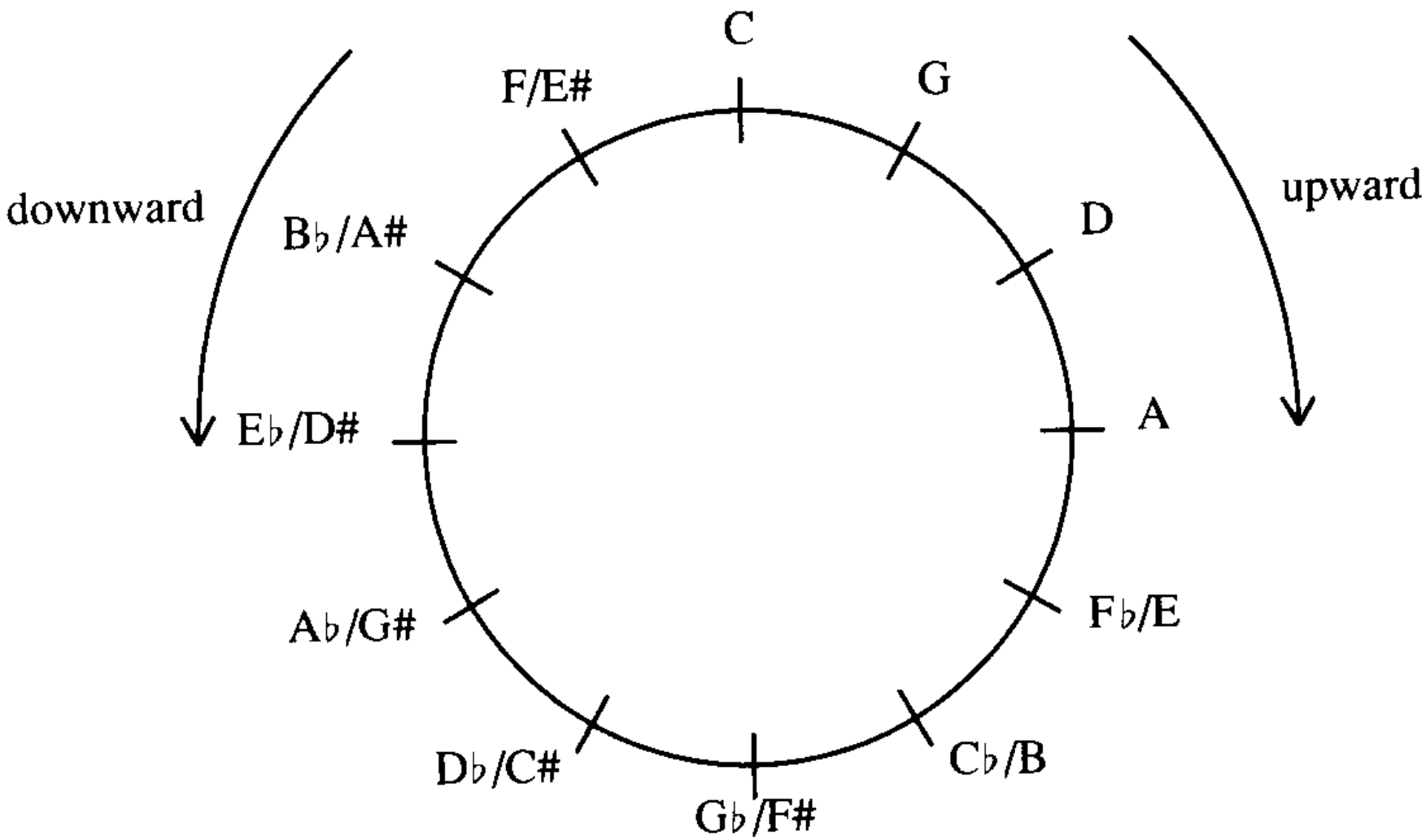


Pythagorean Tuning

- Frequency ratios of all intervals are based on the ratio 3:2, i.e., perfect fifth (P5), which is 7 half-steps.



Circle of Fifths



Problem with Pythagorean Tuning

- One octave = 2f
- A perfect 5^{th} = (3/2)f

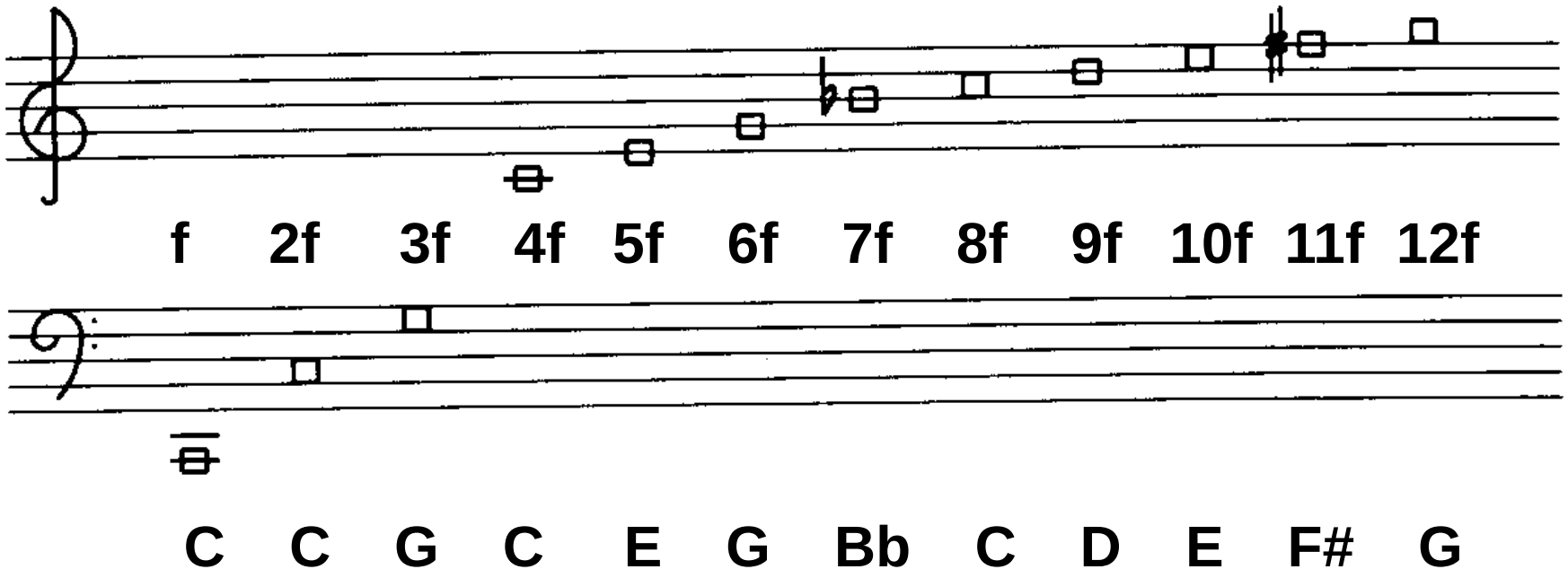
- What happens if you go around the circle of 5ths to get back to your original pitch class?

- $(3/2)^{12}$ = 129.75
- Nearest octave is 2^7 = 128
- 128 != 129.75

- Not convenient for key changes

Overtone Series

- Approximate notated pitch for the harmonics (overtones) of a frequency



Outline

- Basic Concepts of Pitch
- Single Pitch Detection
- Multi-Pitch Analysis

Why is pitch detection important?

- Harmonic sounds are ubiquitous
  - Music, speech, bird singing
- Pitch (F0) is an important attribute of harmonic sounds, and it relates to other properties
  - Music melody → key, scale (e.g., chromatic, diatonic, pentatonic), style, emotion, etc.
  - Speech intonation → word disambiguation (for tonal languages), statement/question, emotion, etc.

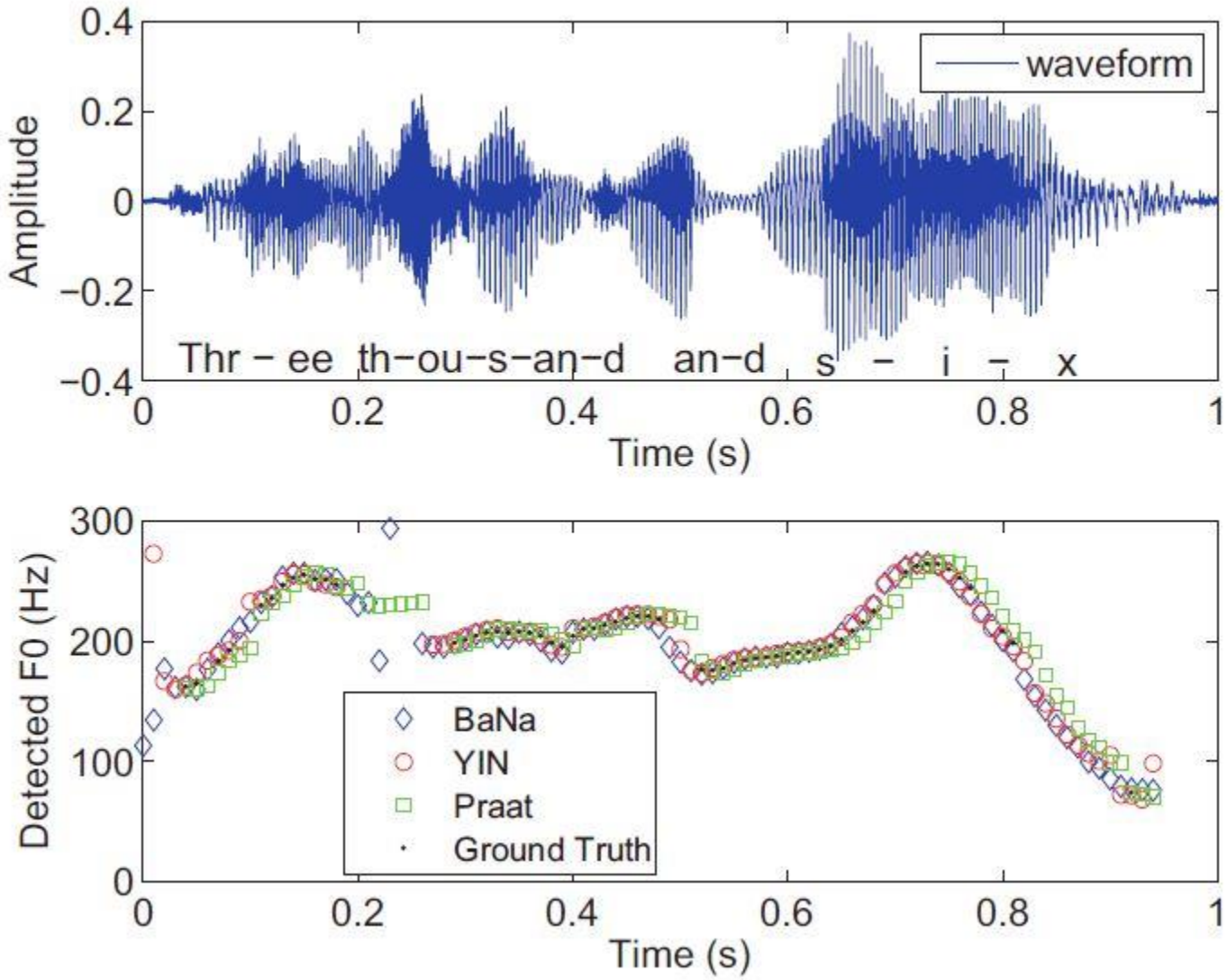
What scales are used?

What emotion?

General Process of Pitch Detection

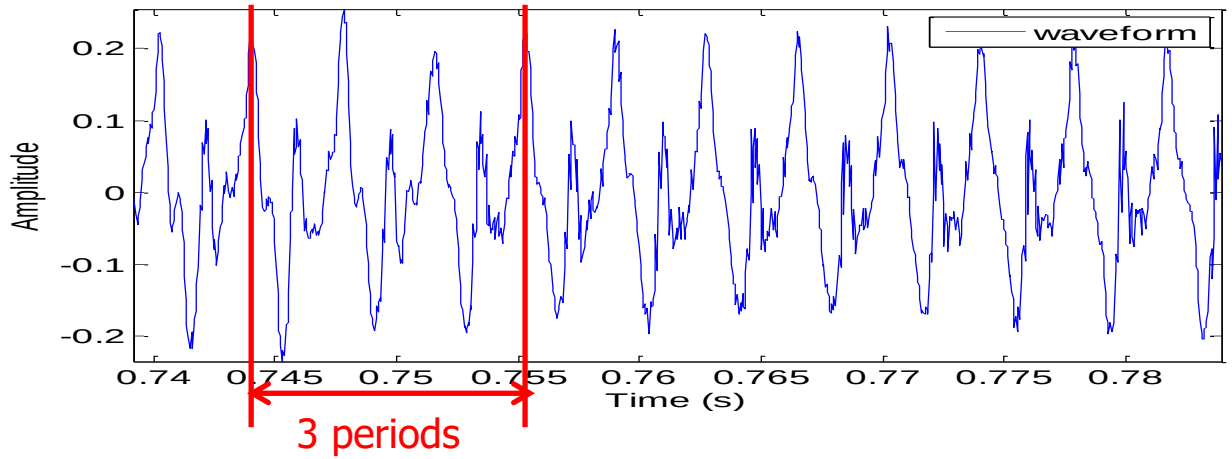
- Segment audio into time frames
  - Pitch changes over time
- Detect pitch (if any) in each frame
  - Need to detect if the frame contains pitch or not
- Post-processing to consider contextual info
  - Pitch contours are often continuous

An Example


waveform
Amplitude
0.4
0.2
0
−0.2
−0.4
Thr − ee th−ou−s−an−d an−d s − i − x
0
0.2
0.4
0.6
0.8
1
Time (s)
Detected F0 (Hz)
300
200
100
0
BaNa
YIN
Praat
Ground Truth
0
0.2
0.4
0.6
0.8
1
Time (s)

How long should the frame be?

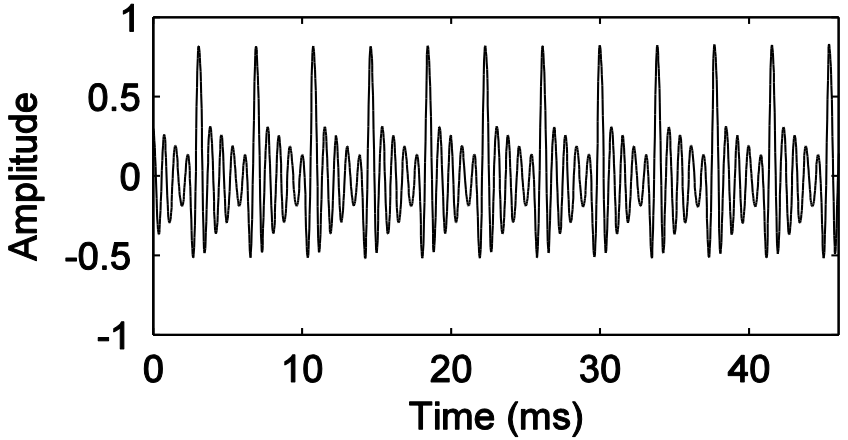
- Too long:
  - Contains multiple pitches (low time resolution)
- Too short
  - Can't obtain reliable detection (low freq resolution)
  - Should be at least about 3 periods of the signal

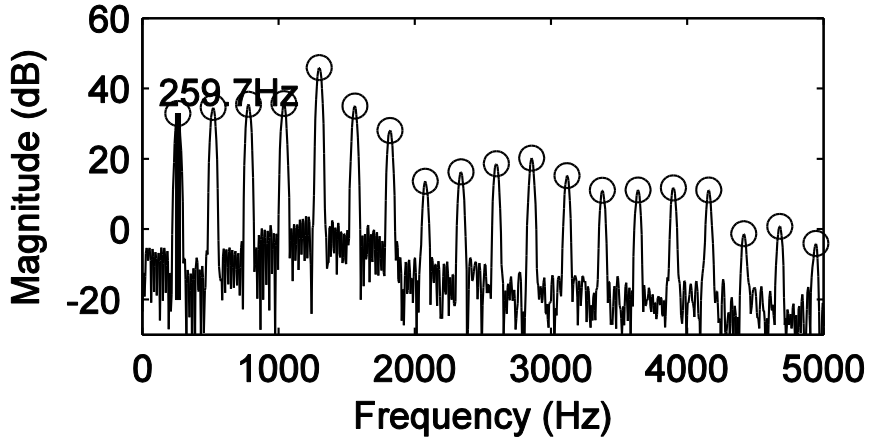


  - For speech or music, how long should the frame be?

Pitch-Related Properties

- Time domain signal is periodic
  - F0 = 1/period
- Spectral peaks have harmonic relations
  - F0 is the greatest common divisor
- Spectral peaks are equally spaced
  - F0 is the frequency gap





Pitch Detection Methods

- Time domain signal is periodic
  - F0 = 1/period
- Spectral peaks have harmonic relations
  - F0 is the greatest common divisor
- Spectral peaks are equally spaced
  - F0 is the frequency gap

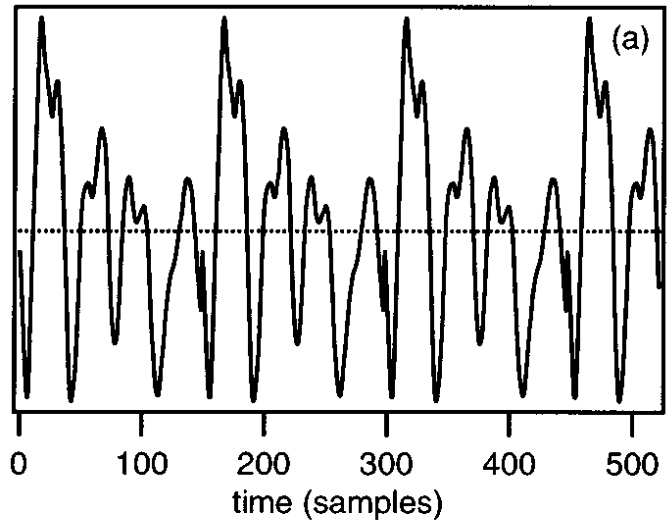
- Time domain
  - Detect period
- Frequency domain
  - Detect the divisor
- Cepstrum domain
  - Detect the gap

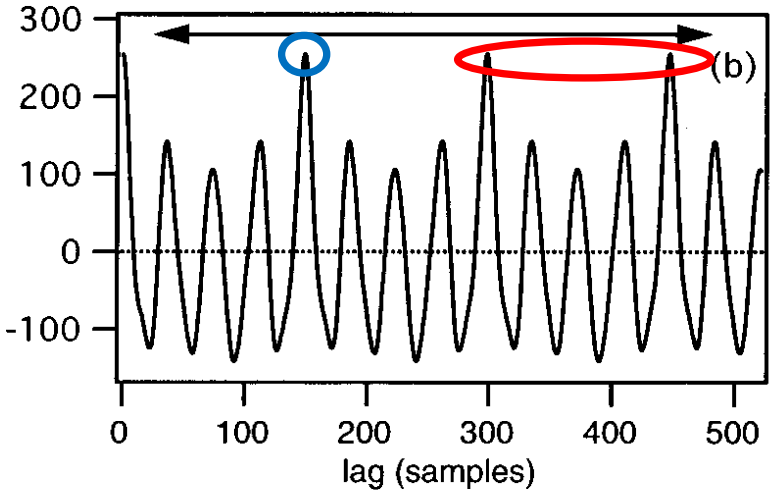
Time Domain: Autocorrelation

$$r_t(\tau) = \sum_{j=t+1}^{t+W} x_j x_{j+\tau}$$

- A periodic signal correlates strongly with itself when offset by the period (and multiple periods)

- Problem: sensitive to peak amplitude changes
  - Which peak would be higher if signal amplitude increases?
  - Lower octave error (or sub-harmonic error)





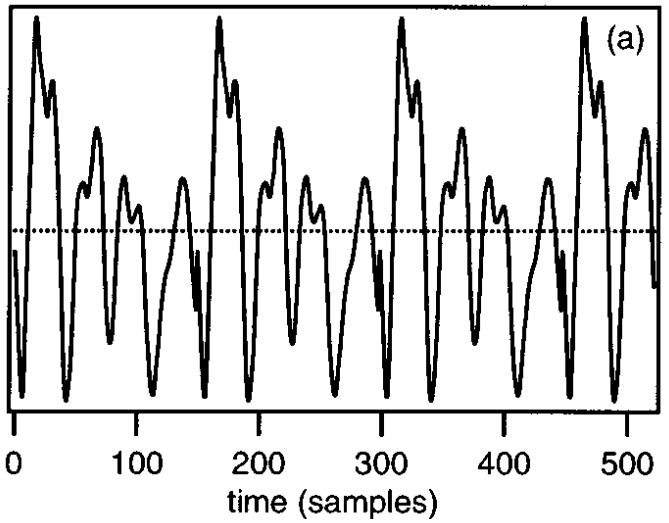
YIN: Autocorrelation → Difference Function

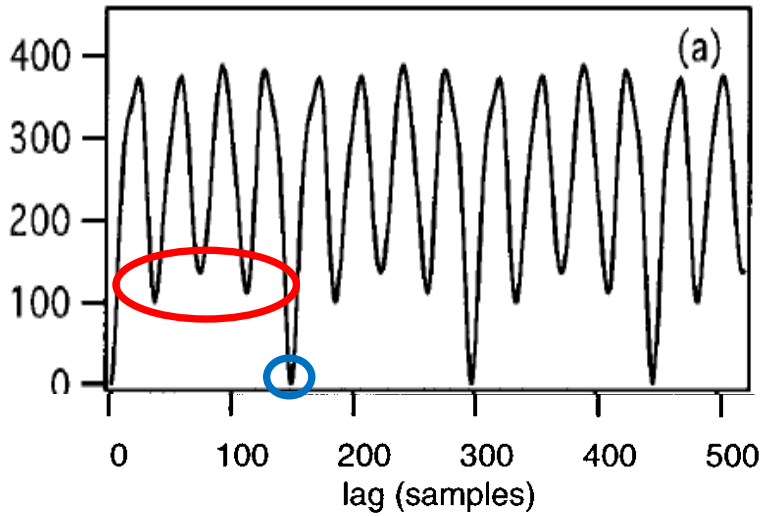
- Replace ACF with difference function

$$d_t(\tau) = \sum_{j=1}^{W} (x_j - x_{j+\tau})^2$$

- Look for dips instead of peaks, which is why it's called YIN opposed to YANG.
- Immune to amplitude changes
- Problem
  - Some dips close to 0 lag might be deeper due to imperfect periodicity
- YIN algorithm has several other steps to fix this and other issues

[de Cheveigne, 2002]



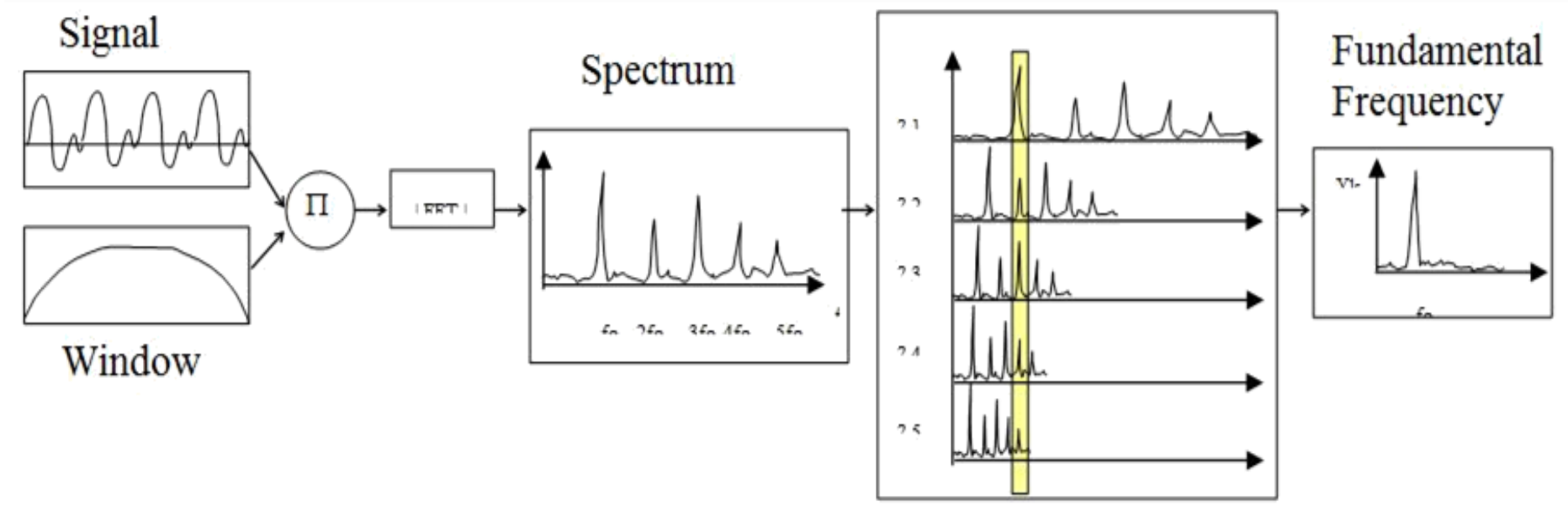


Frequency Domain Approach

- Idea: for each F0 candidate, calculate the support (e.g., spectral energy) it receives from its harmonic positions.

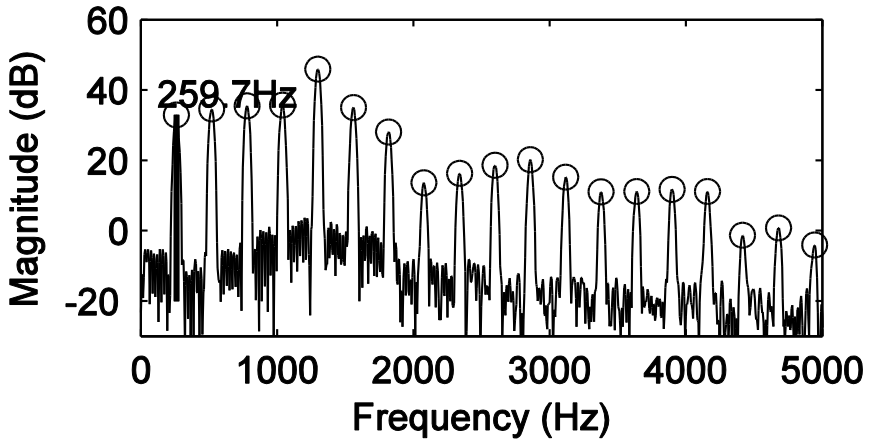
- E.g., Harmonic Product Spectrum (HPS)

[Schroeder, 1968; Noll, 1970]



Cepstral Domain Approach

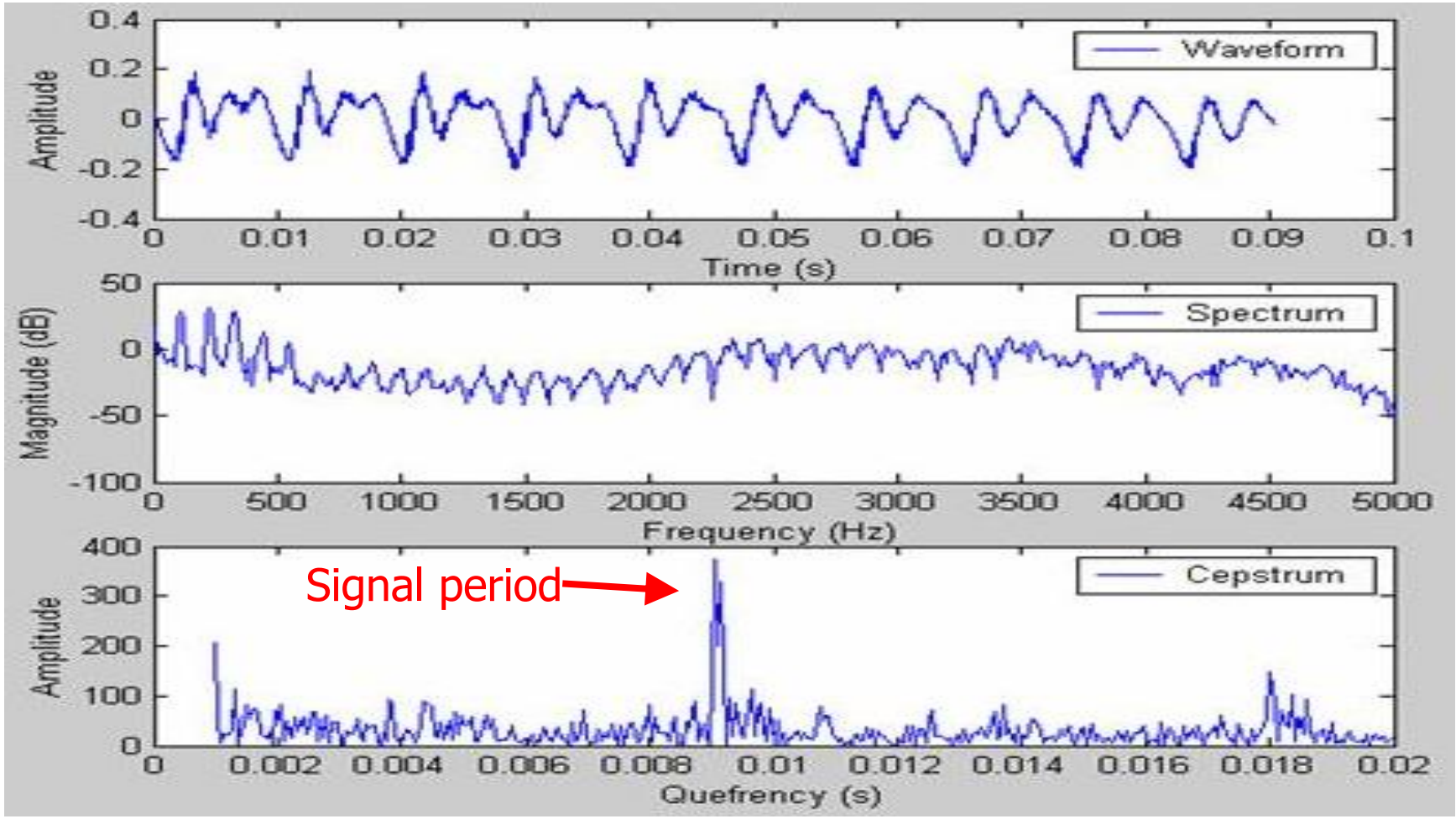
- Idea: find the frequency gap between adjacent spectral peaks
  - The log-amplitude spectrum looks pretty periodic
  - The gap can be viewed as the period of the spectrum
  - How to find the period then?
  - Cepstrum idea: Fourier transform!



Cepstrum

$$\text{power cepstrum} = |\mathcal{F}^{-1}\{\log|\mathcal{F}\{x(t)\}|^2\}|^2$$

Spectrum - Cepstrum
Frequency - Quefrency
Filtering - Liftering



Pitched or Non-pitched?

- Some frames may be silent or inharmonic, so they may not contain a pitch at all
  - Silence can be detected by RMS value
  - How about inharmonic frames?

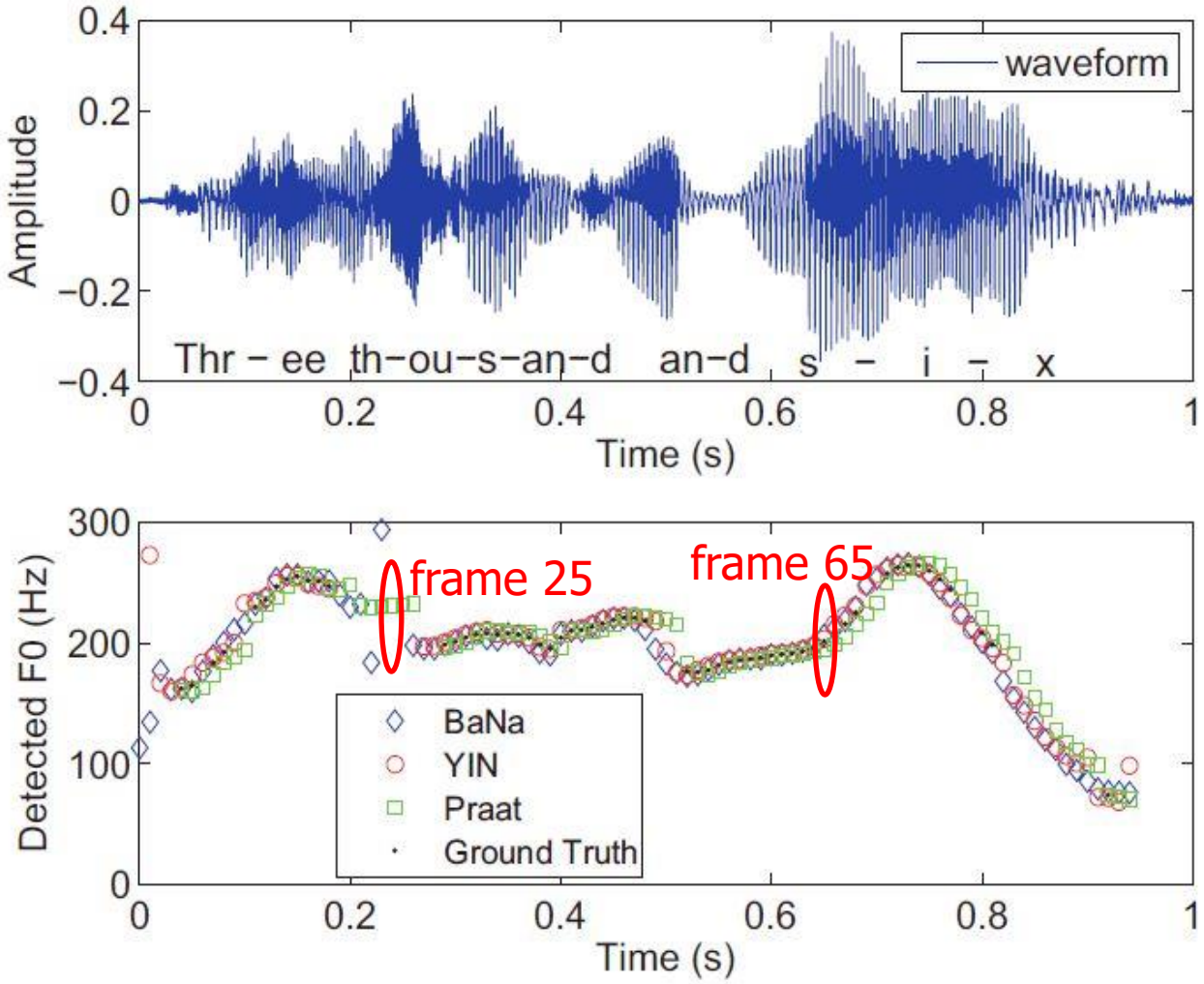
- YIN: threshold on dip, aperiodicity
- HPS: threshold on the peak amplitude of the product spectrum
- Cepstrum: threshold on ratio between amplitudes of the two highest cepstral peaks
  - [Rabiner 1976]

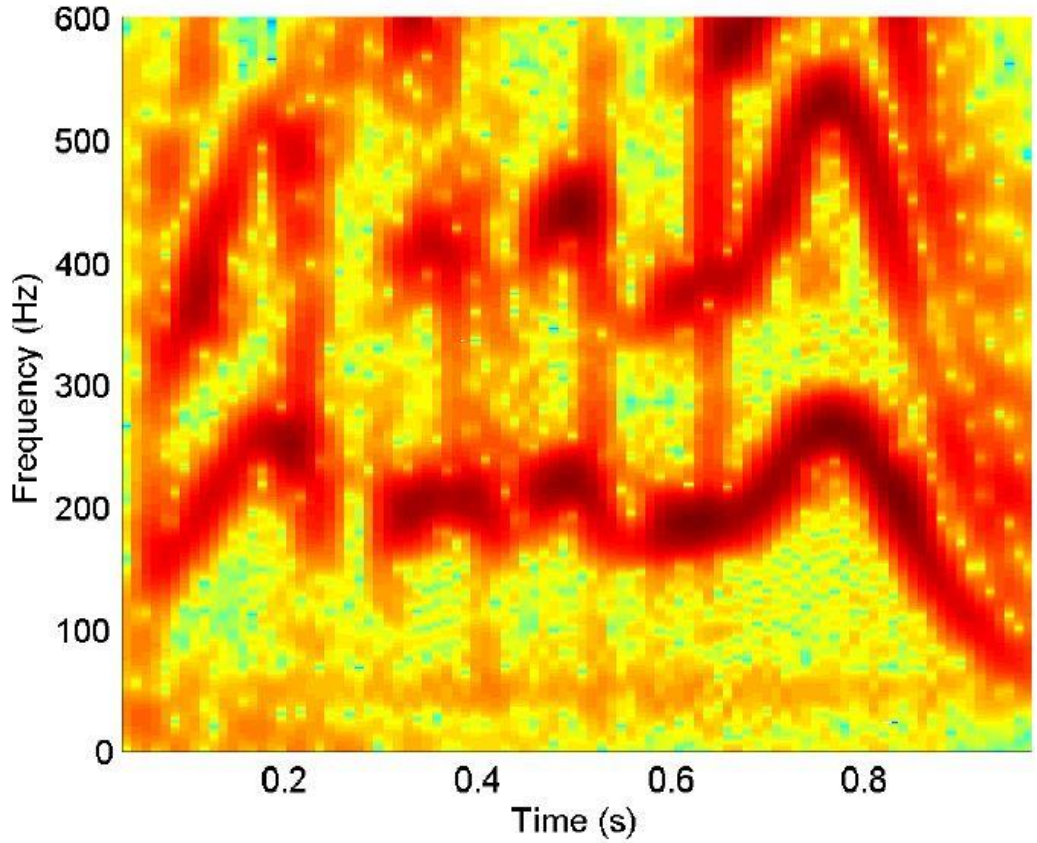
How to evaluate pitch detection?

- Choose some recordings (speech, music)
- Get ground-truth

- Pitched/non-pitched classification error
- Calculate the difference between estimated pitch with ground-truth
  - Threshold for speech: 10% or 20% in Hz
  - Threshold for music: 1 quarter-tone (about 3% in Hz)

Different Methods vs. Ground-truth





He Ba, Na Yang, I. Demirkol and W. Heinzelman, "BaNa: A hybrid approach for noise resilient pitch detection," *IEEE Statistical Signal Processing Workshop (SSP)*, 2012

Pitch Detection with Noise

- Can we still hear pitch if there is some background noise, say in a restaurant?

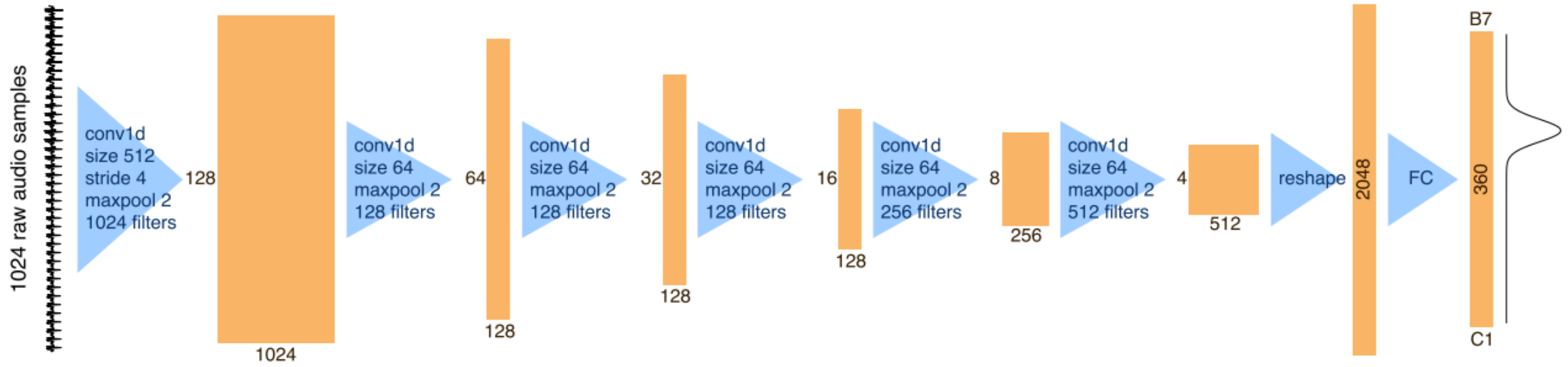
Violin + babble noise

- Will pitch detection algorithms still work?
- Which domain is less sensitive to which kind of noise?
- How to improve pitch detection in noisy environments?

Supervised Learning Method

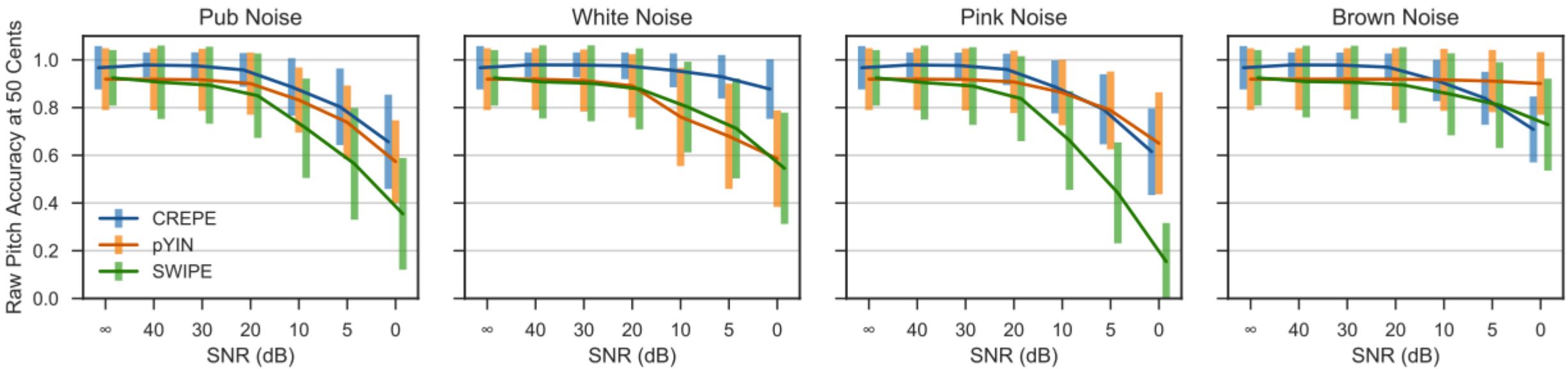
- Model input: audio frame; model output: target pitch
- Data driven; can be trained on specific type of data or diverse data

- CREPE: Convolutional Representation for Pitch Estimation



Kim, Jong Wook, et al. "CREPE: A convolutional representation for pitch estimation." *In Proc. ICASSP 2018*.

More Robust to Noise

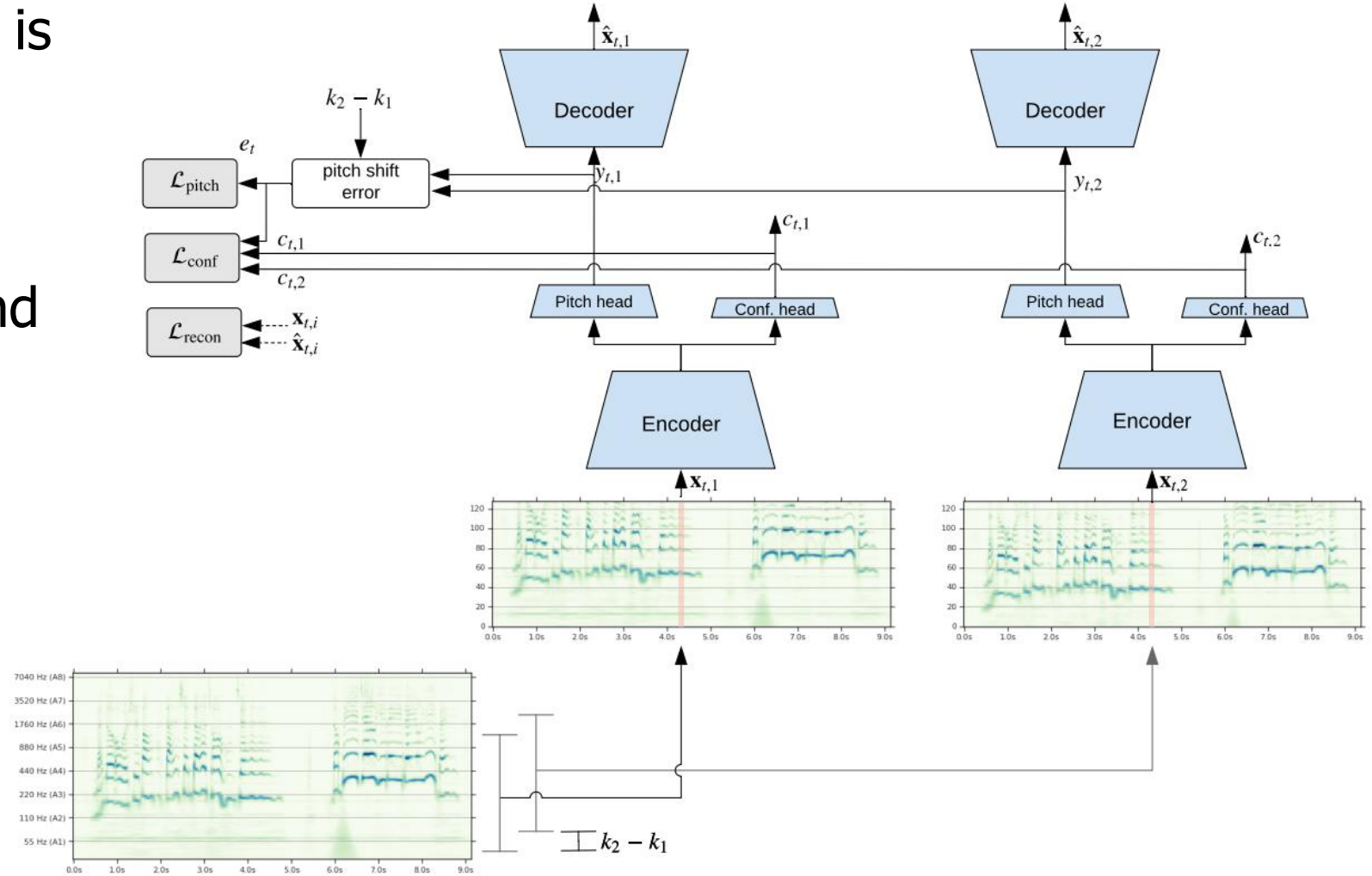


- Robustness can be further improved with data augmentation
- Online repo and model: https://github.com/marl/crepe

- Limitation: Requires annotated data to train

SPICE: Self-supervised PItCh Estimation

- Inspiration: relative pitch is easier to transcribe than absolute pitch
- Training: Feeding CQT spectrograms (original and transposed)
- Calibration: using a small synthetic dataset to get absolute pitch



Gfeller, Beat, et al. "SPICE: Self-supervised pitch estimation." *IEEE/ACM TASLP 2020.*

Result Comparisons

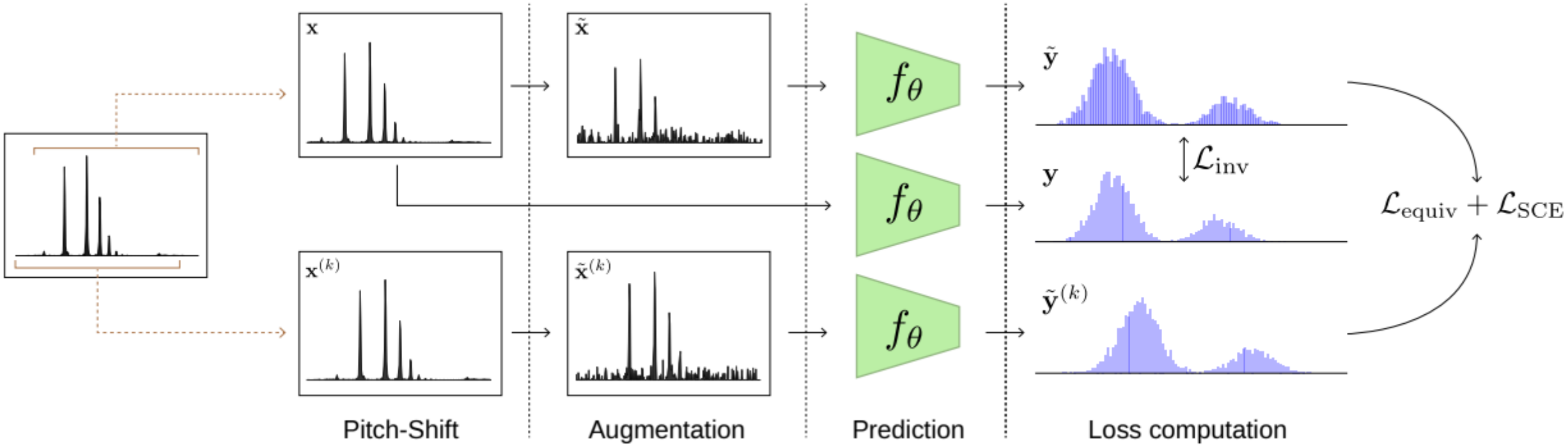
Model	# params	Trained on	*MIR-1k* RPA (CI 95%)	*MIR-1k* VRR	*MDB-stem-synth* RPA (CI 95%)
SWIPE	-	-	86.6%	-	90.7%
CREPE tiny	487k	many	90.7%	88.9%	93.1%
CREPE full	22.2M	many	90.1%	84.6%	92.7%
SPICE	2.38M	*SingingVoices*	$90.6\% \pm 0.1\%$	86.8%	$89.1\% \pm 0.4\%$
SPICE	180k	*SingingVoices*	$90.4\% \pm 0.1\%$	90.5%	$87.9\% \pm 0.9\%$

Model	# params	Trained on	*MIR-1k* clean	*MIR-1k* 20dB	*MIR-1k* 10dB	*MIR-1k* 0dB
SWIPE	-	-	86.6%	84.3%	69.5%	27.2%
CREPE tiny	487k	many	90.7%	90.6%	88.8%	76.1%
CREPE full	22.2M	many	90.1%	90.4%	89.7%	80.8%
SPICE	2.38M	*MIR-1k* + augm.	$91.4\% \pm 0.1\%$	$91.2\% \pm 0.1\%$	$90.0\% \pm 0.1\%$	$81.6\% \pm 0.6\%$

- Smaller model
- Decent performance
- Better noise robustness

PESTO

Alain Riou, Stefan Lattner, Gaëtan Hadjeres, Geoffroy Peeters. "PESTO: Pitch Estimation with Selfsupervised Transposition-equivariant Objective," in *Proc. ISMIR*, 2023 (best paper award!)



- Training input: CQT spectrum + its transposed and augmented versions
- Training target: invariance, equivariance, shifted cross entropy

Result Comparisons

Model	# params	Trained on	Raw Pitch Accuracy	
			MIR-1K	*MDB-stem-synth*
SPICE [19]	2.38M	private data	90.6%	89.1%
DDSP-inv [45]	-	*MIR-1K* / *MDB-stem-synth*	91.8%	88.5%
PESTO (ours)	28.9k	*MIR-1K*	**96.1%**	94.6%
PESTO (ours)	28.9k	*MDB-stem-synth*	93.5%	**95.5%**
CREPE [16]	22.2M	many (supervised)	**97.8%**	**96.7%**

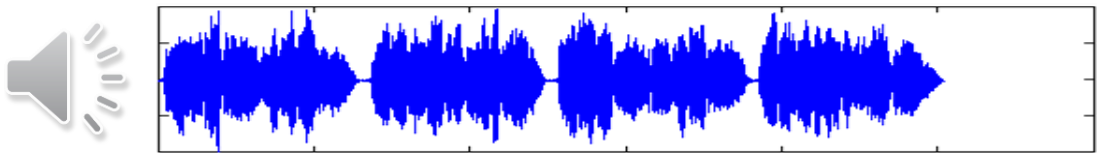
- Extremely light model
- Comparable results to supervised method

Outline

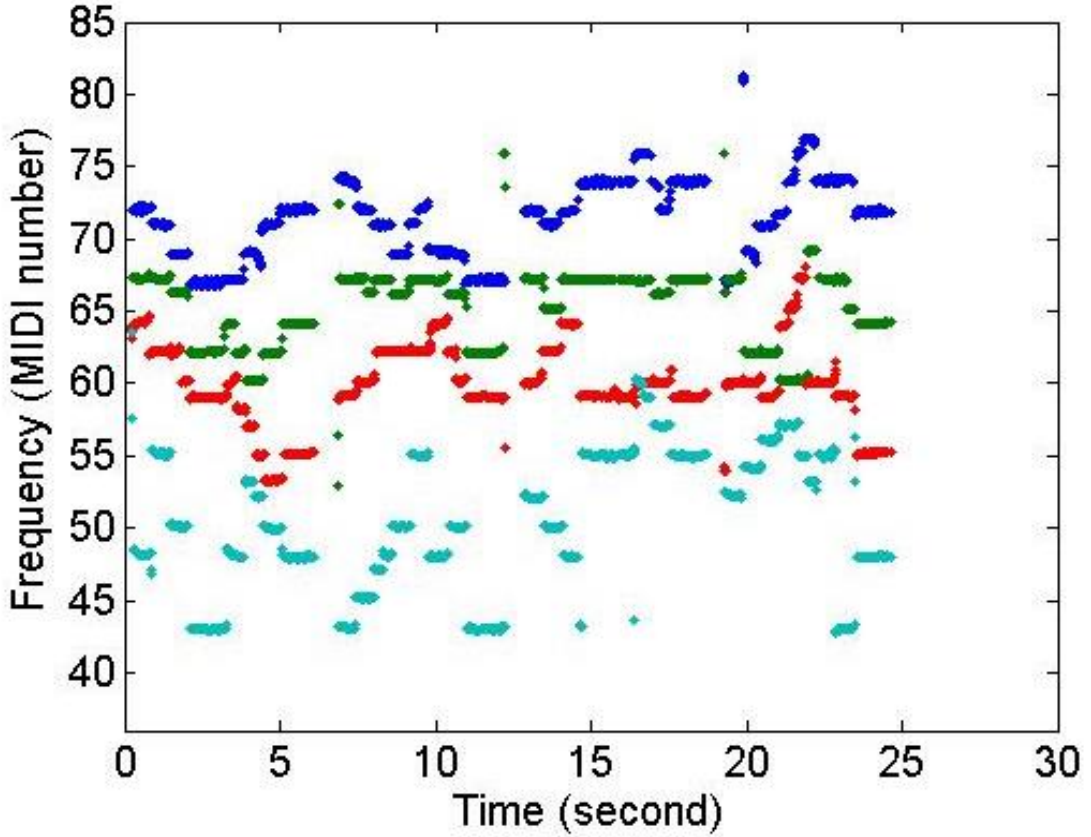
- Basic Concepts of Pitch
- Single Pitch Detection
- Multi-Pitch Analysis
  - Many slides are copied from ISMIR 2015 Tutorial on "Automatic Music Transcription", which provides a much more comprehensive review: https://c4dm.eecs.qmul.ac.uk/ismir15-amt-tutorial/

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
  - ……

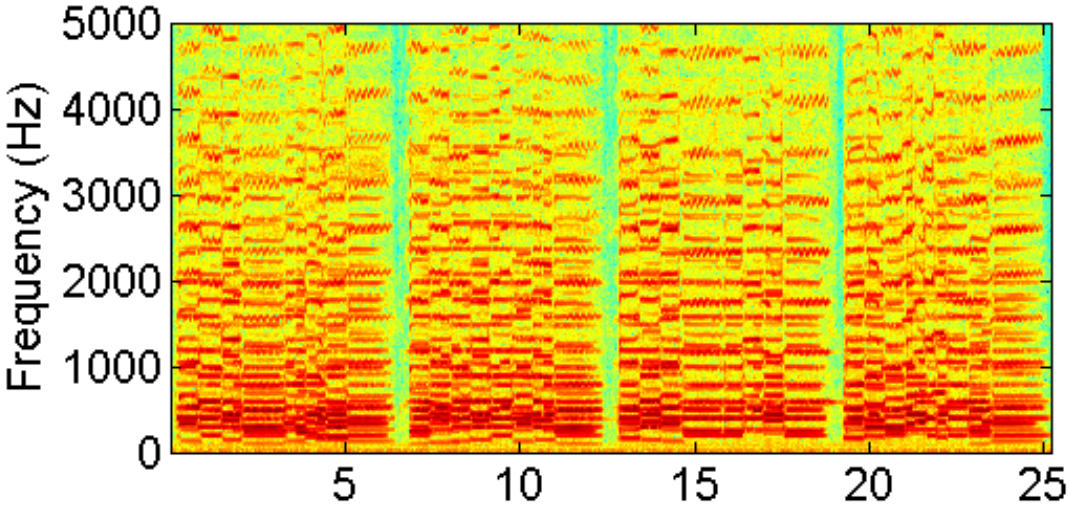
How difficult is it?

- 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?

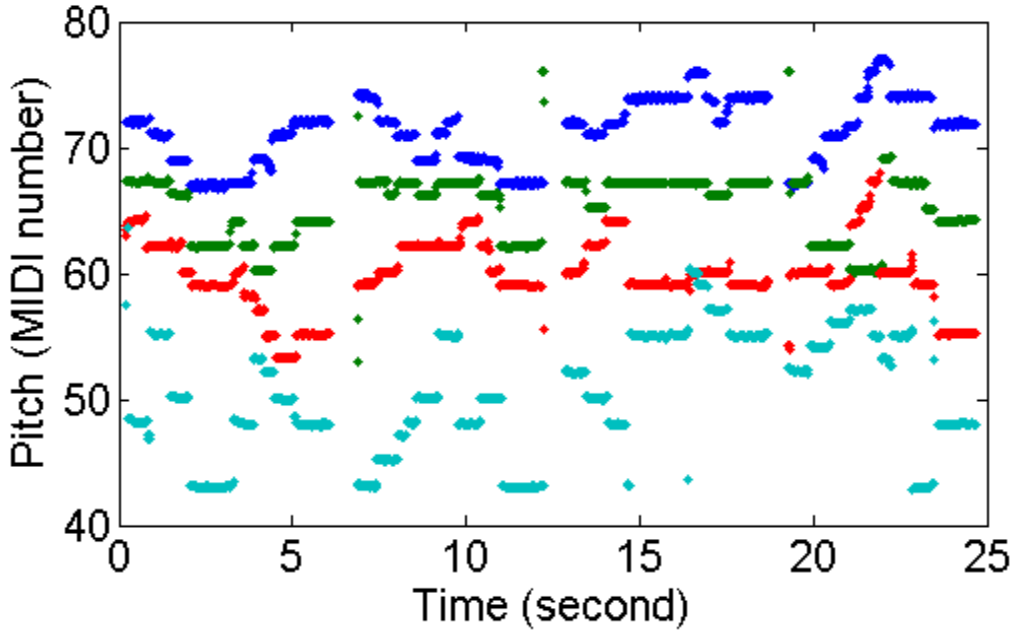
Chord 1	Chord 2
2	3
C4/G4	C4/F4/A4
Clarinet G4 Horn C4	Clarinet A4 Viola F4 Horn C4

Our Task

Spectrogram

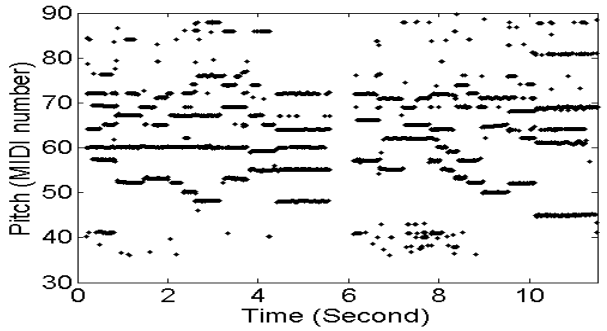


Ground-truth pitch trajectories

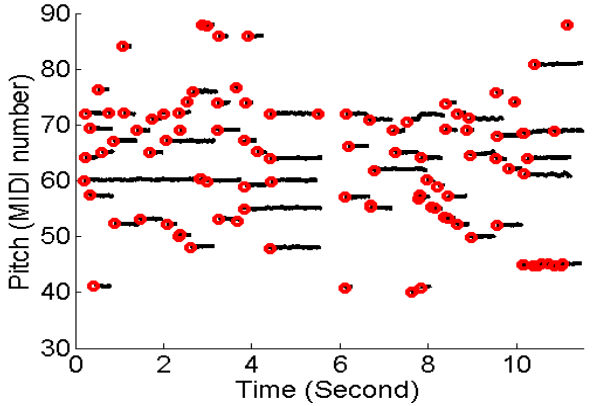


Three Levels of Multi-pitch Analysis

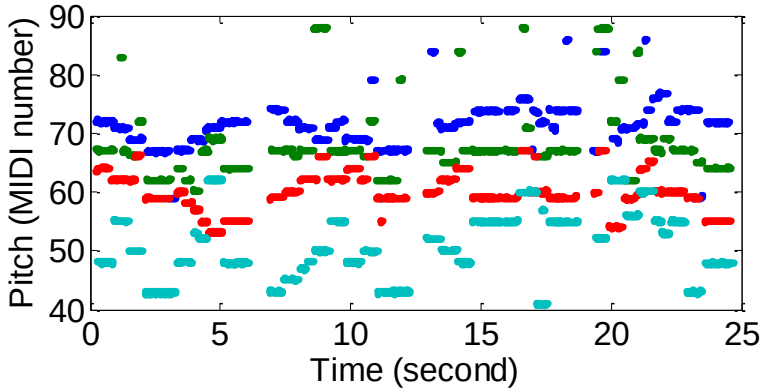
- Frame-level (multi-pitch estimation)
  - Estimate pitches and polyphony in each frame
  - Many methods



- Note-level (note tracking)
  - Estimate pitch, onset, offset of notes
  - Fewer methods

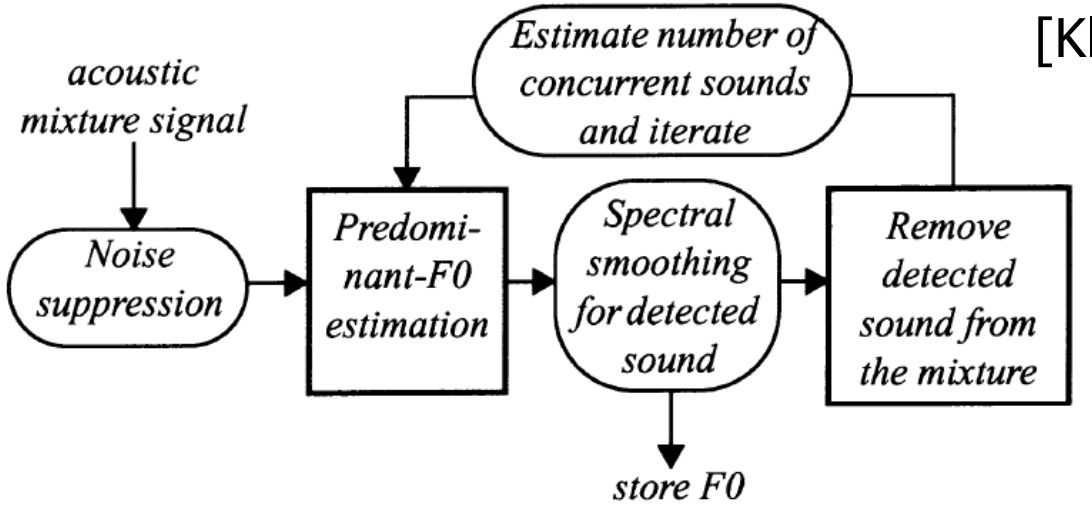


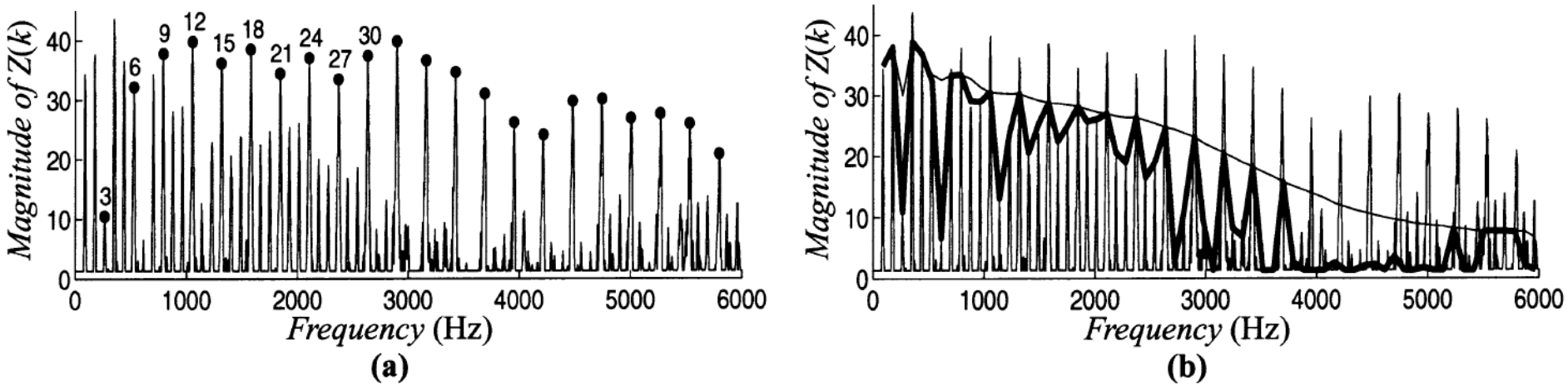
- Stream-level (multi-pitch streaming)
  - Stream pitches by sources
  - Very few methods



Iterative Spectral Subtraction

[Klapuri, 2003]





Pros: good performance, simple, fast

Cons: hard to subtract the appropriate amount of energy

Spectral Peak Modeling – Maximum Likelihood

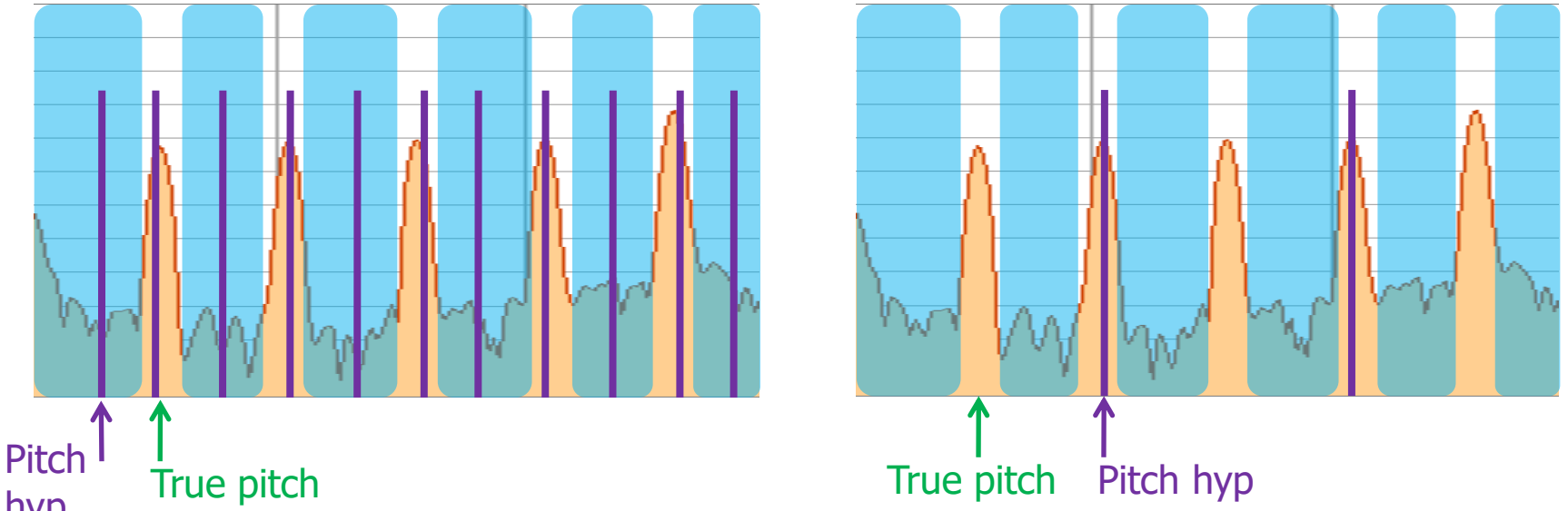
- [Duan et al., 2010]

Pros: balances harmonic and subharmonic errors
Cons: soft notes may be masked by others

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

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 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

Full Spectrum Modeling – Probabilistic

- Key idea: view spectra as (parametric) probabilistic distributions

- Each note = tied- Gaussian Mixture Model (tied-GMM)

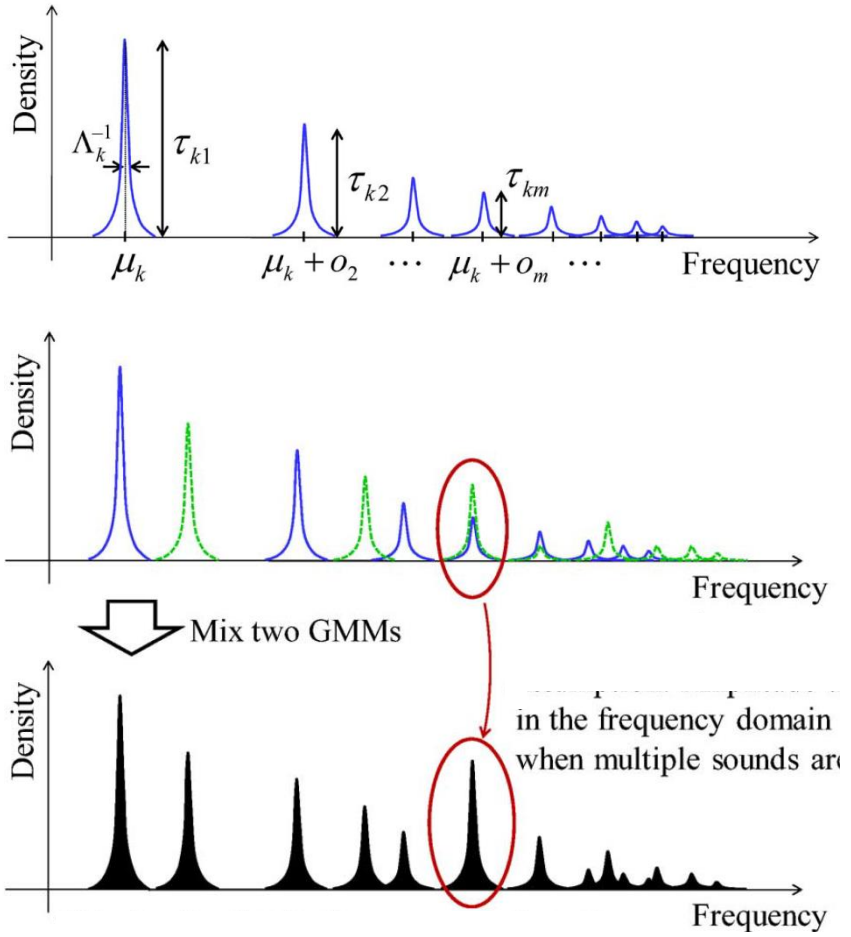
$$\mathcal{M}_k(\boldsymbol{x}) = \sum_{m=1}^{M} \tau_{km} \mathcal{N}\left(\boldsymbol{x} \middle| \boldsymbol{\mu}_k + \boldsymbol{o}_m, \boldsymbol{\Lambda}_k^{-1}\right)$$

- Signal = Mixture of GMMs

$$\mathcal{M}_d(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_{dk} \mathcal{M}_k(\boldsymbol{x})$$

Pros: flexible to incorporate priors on parameters

Cons: doesn't model inharmonic and transients; many parameters to optimize



Figures from [Yoshii & Goto, 2012]

Full Spectrum Modeling – Probabilistic

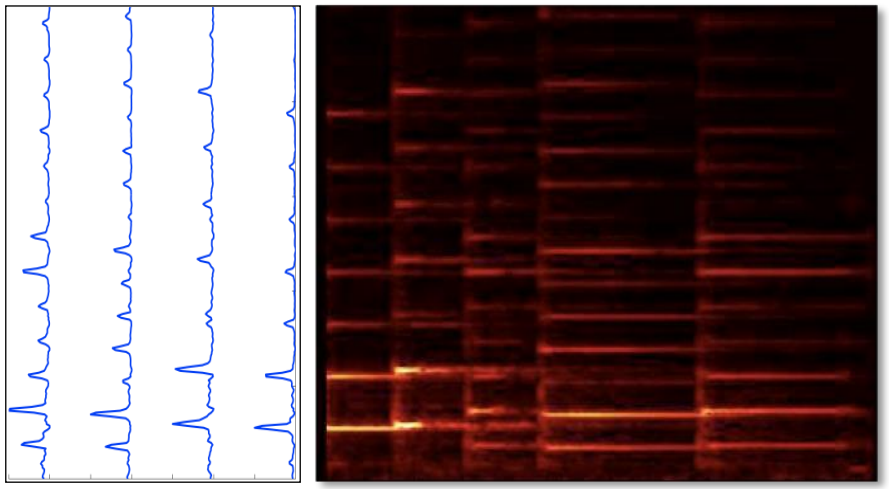
Non-parametric model

[Smaragdis & Raj, 2006]

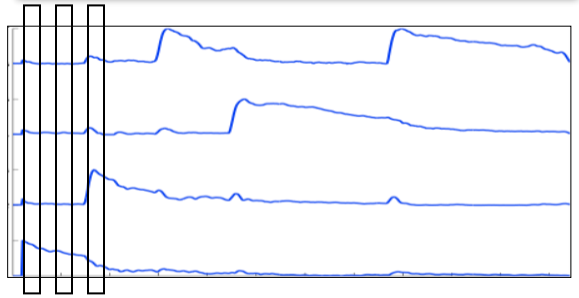
- Probabilistic Latent Component Analysis (PLCA)

Sound quanta distribution at t → $P_t(f) \approx \sum_z P(f|z)P_t(z)$

Dictionary Elements $P(f|z)$ →



Time-invariant sound quanta distribution for each component



← Activation weights $P_t(z)$

Distribution of components

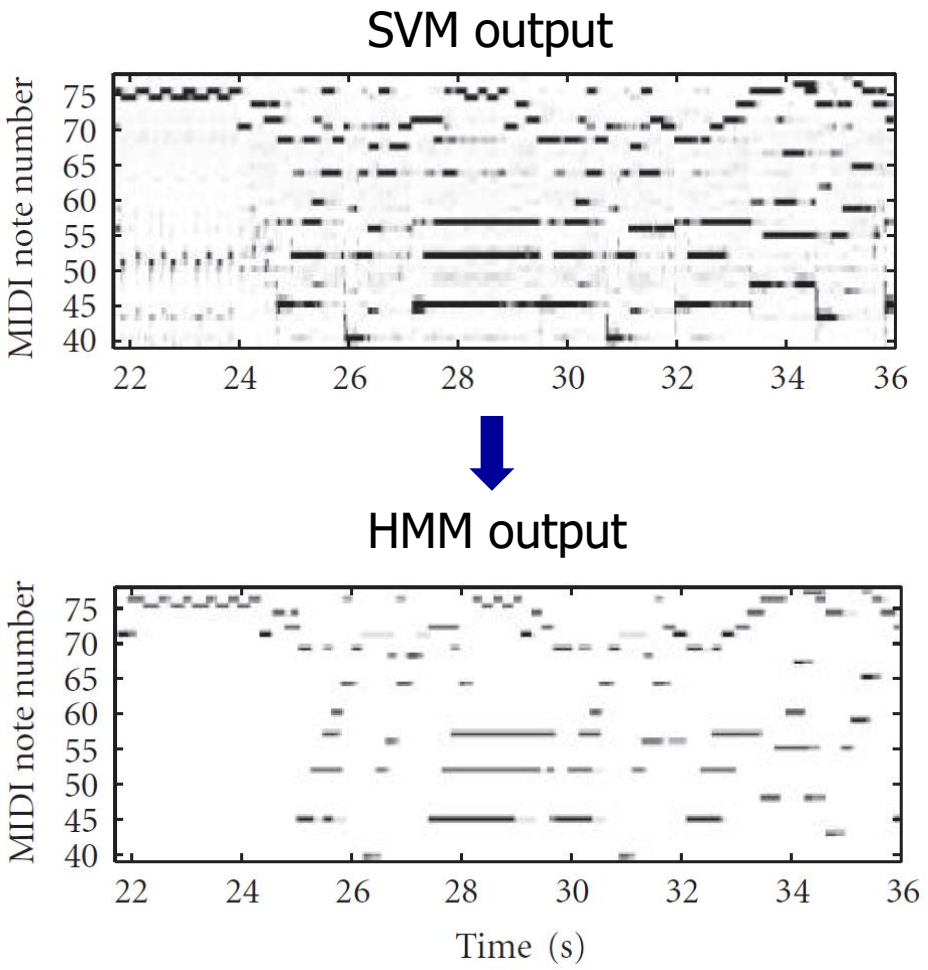
Classification-based Piano Transcription

[Poliner & Ellis, 2007]

- 87 independent one-vs-all SVMs for piano (except for the highest note C8)
- Trained on MIDI-synthesized piano performances
- Features: magnitude spectrum within

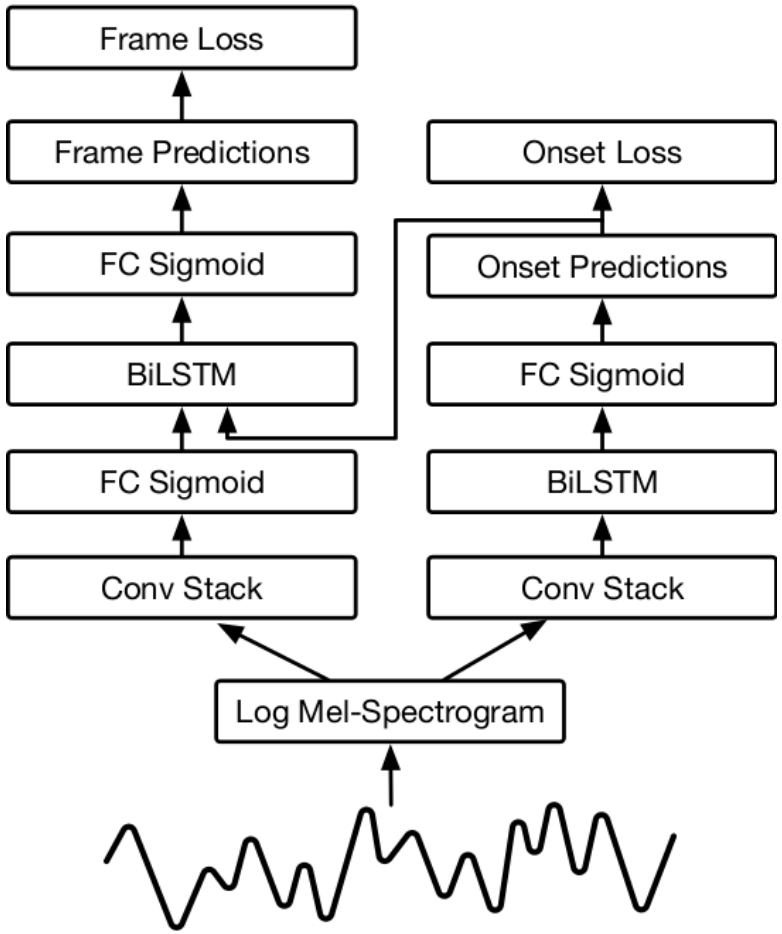
$$\begin{cases} \text{0–2 kHz, for notes} \leq \text{B5 (988Hz)} \\ \text{1–3 kHz, for C6} \leq \text{notes} \leq \text{B6} \\ \text{2–4 kHz, for notes} \geq \text{C7 (2093Hz)} \end{cases}$$

- HMM smoothing for each class independently



Classification-based Piano Note Transcription

[Harthorne et al., 2018]

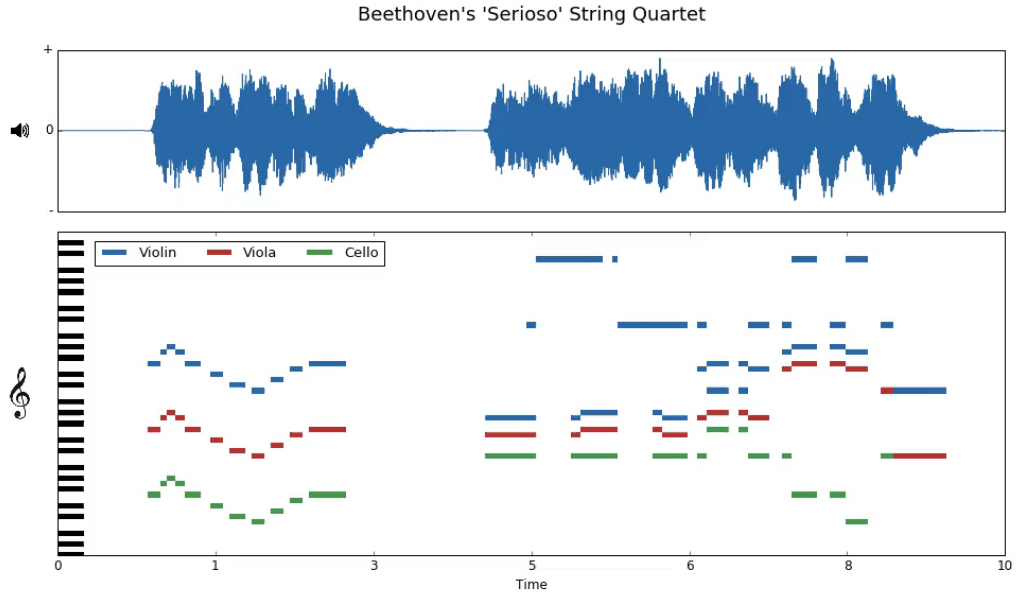


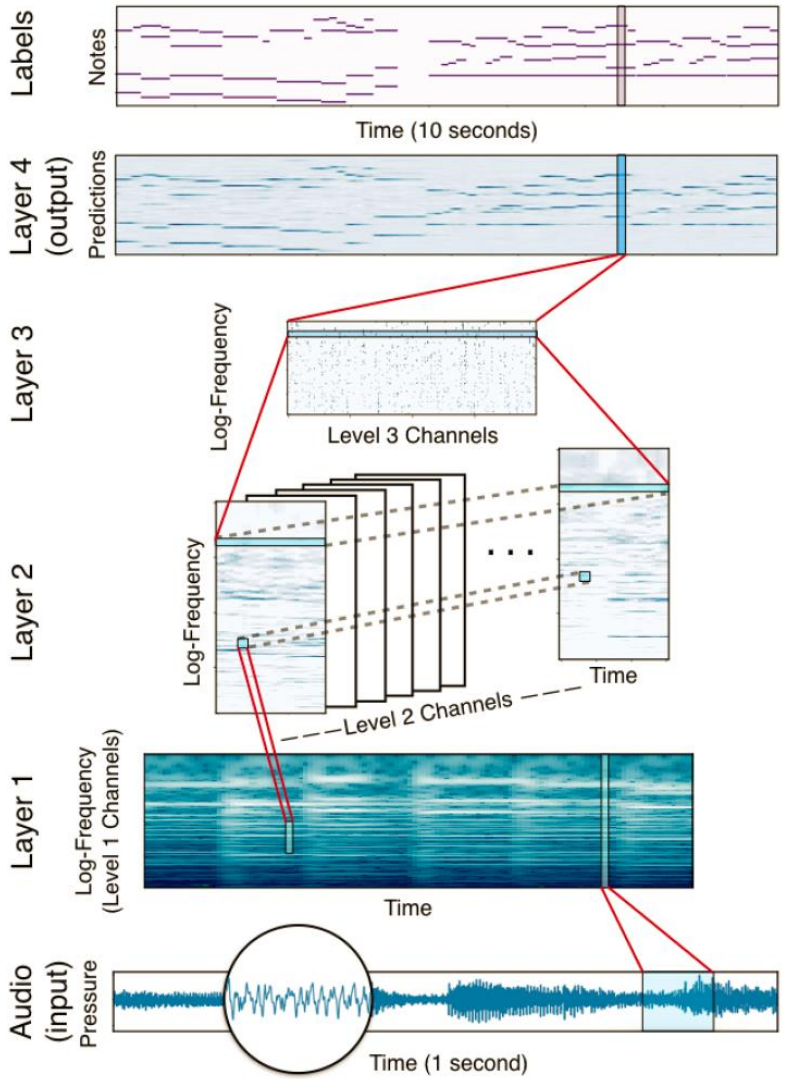
$$L_{frame} = \sum_{p=p_{min}}^{p_{max}} \sum_{t=0}^{T} CE\left(\mathbf{I}_{frame}(p,t), \mathbf{P}_{frame}(p,t)\right)$$

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

Multi-Instrument Transcription

- MusicNet [1]
  - 330 classical pieces with MIDI alignments using Dynamic Time Warping (DTW)

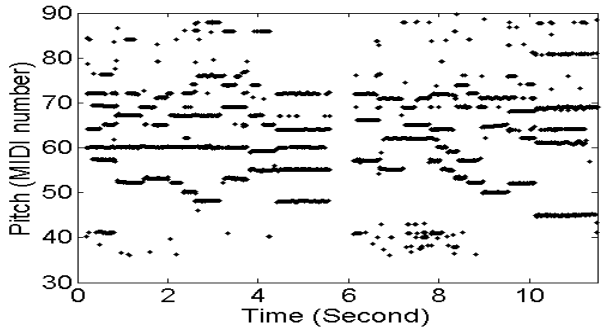




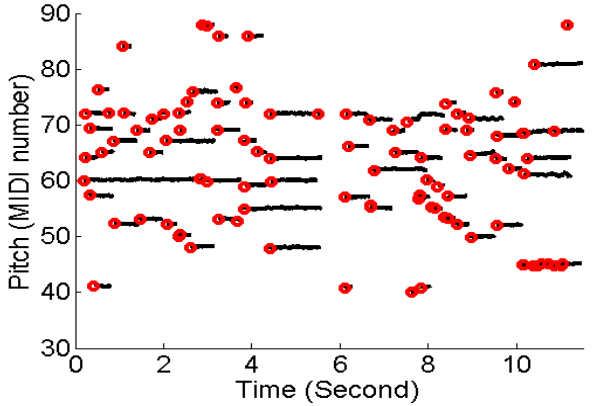
[1] J. Thickstun, Z. Harchaoui, and S. Kakade, **Learning features of music from scratch**, ICLR, 2017.
[2] J. Thickstun, Z. Harchaoui, D.P. Foster, S.M. Kakade, **Invariances and data augmentation for supervised music transcription,** ICASSP, 2018**.**

State of the Art of Multi-pitch Analysis

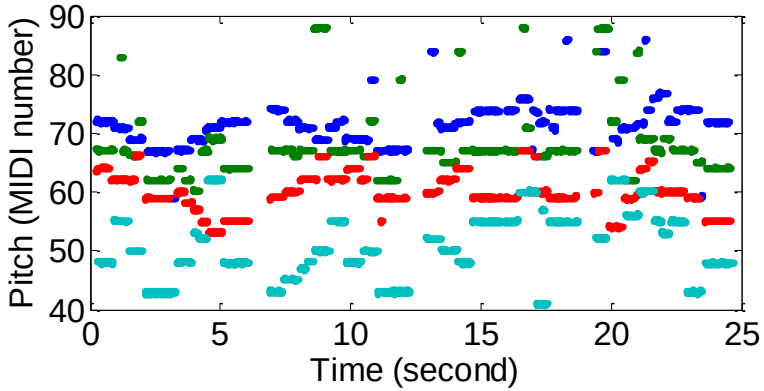
- Frame-level (multi-pitch estimation)
  - Estimate pitches and polyphony in each frame
  - Many methods



- Note-level (note tracking)
  - Estimate pitch, onset, offset of notes
  - Fewer methods



- Stream-level (multi-pitch streaming)
  - Stream pitches by sources
  - Very few methods



Frame Level → Note Level

- Based on pitch salience/likelihood/activations
  - Thresholding, filling, pruning
  - Median filtering: [Su & Yang, 2015]
  - Pitch-wise on/off HMMs

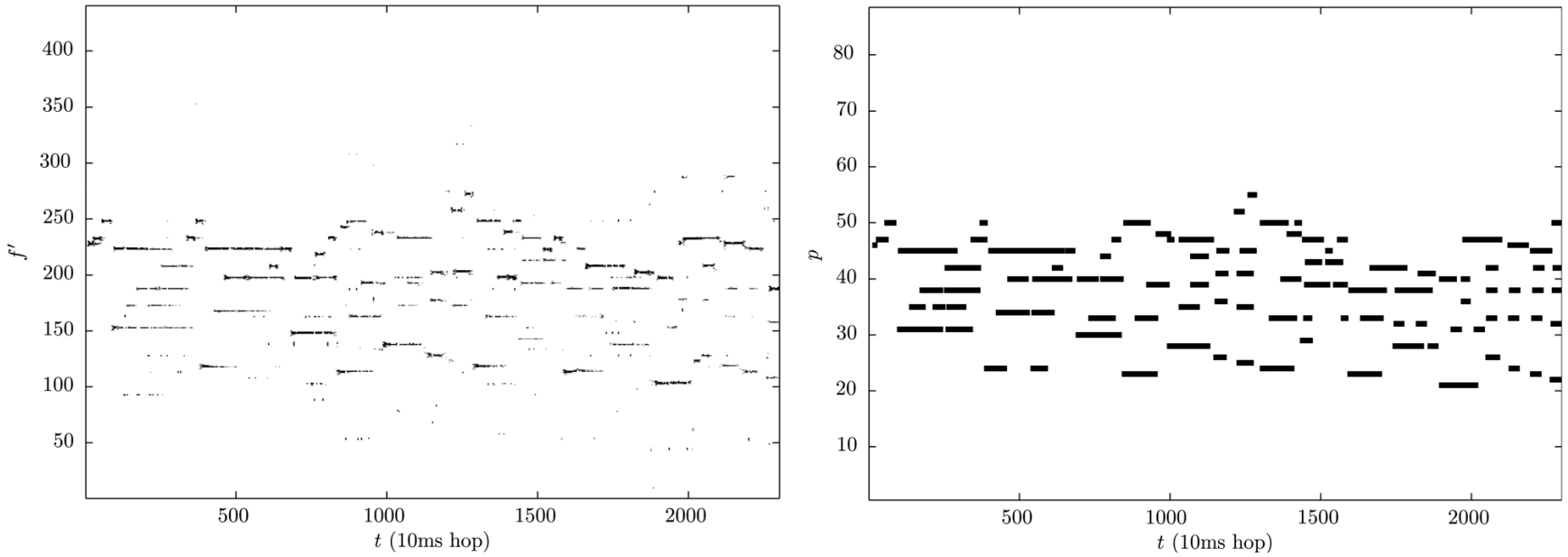


Figure from [Benetos & Dixon, 2013]

Note Tracking from Audio Directly

[Kameoka et al., 2007]

- Harmonic temporal structured clustering (HTC)

Mixture spectrogram

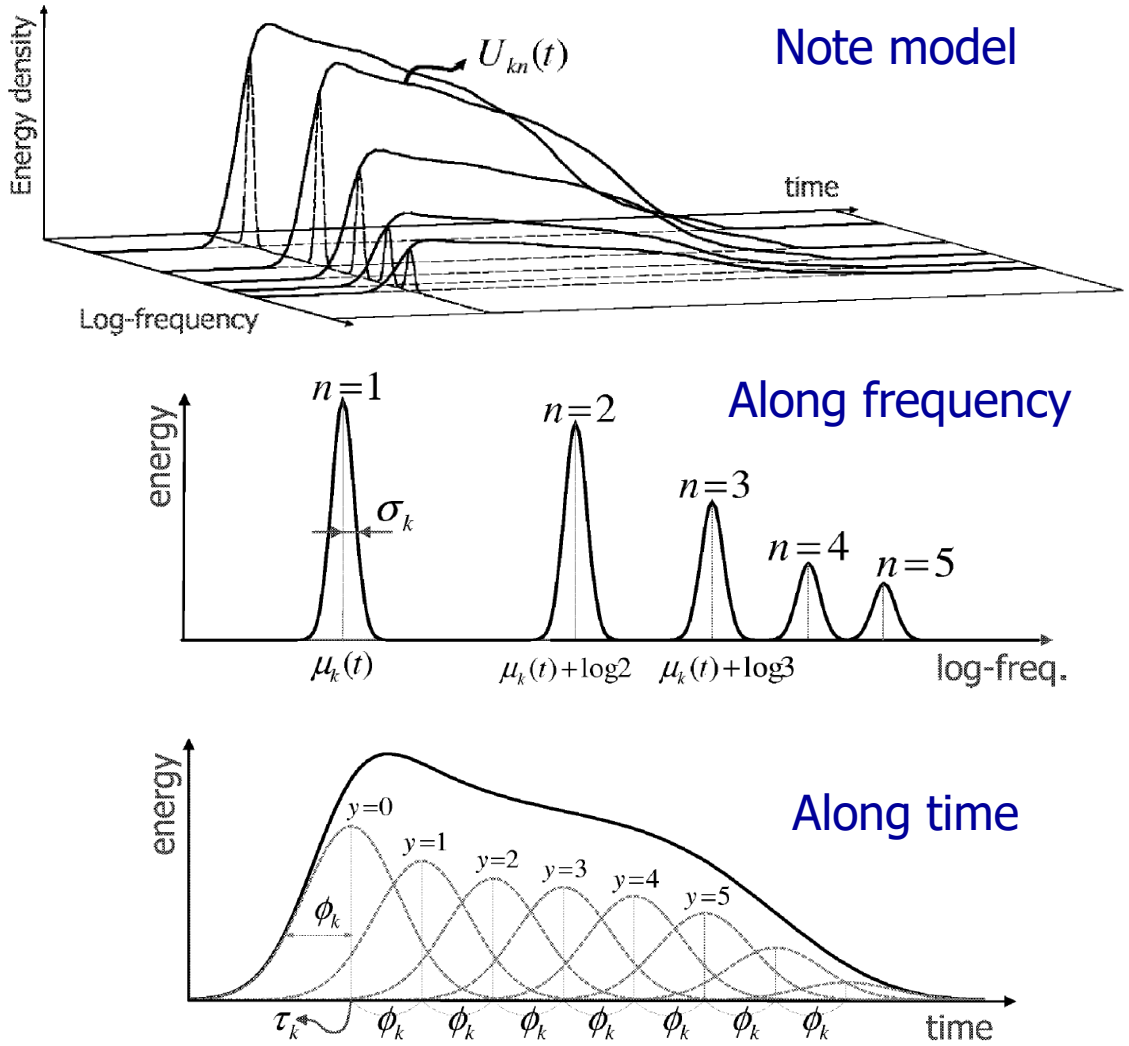
Activation of sources (latent variables)

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

Source signal

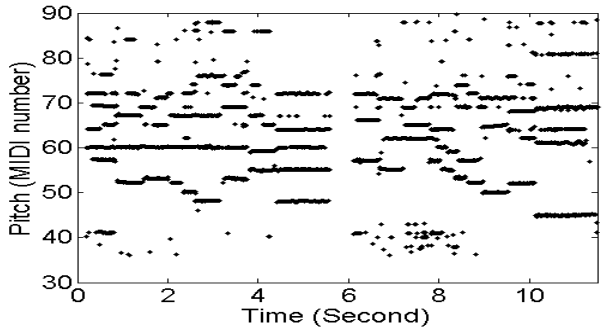
parameters

- EM algorithm

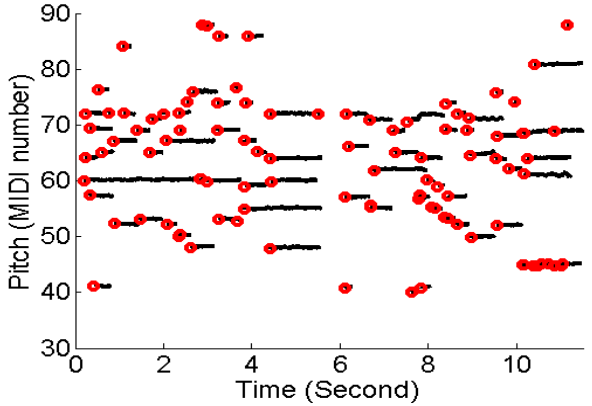


State of the Art

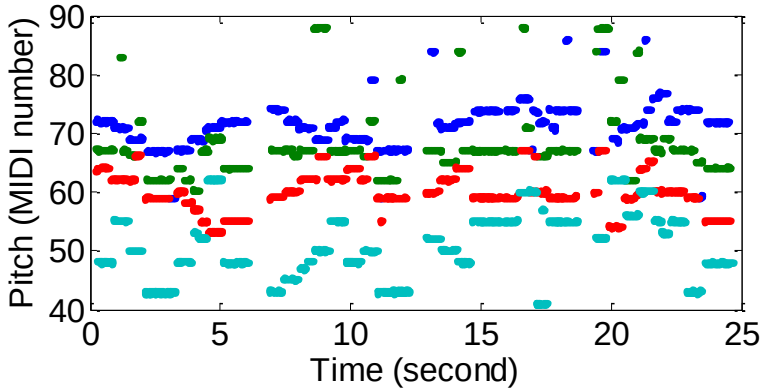
- Frame-level (multi-pitch estimation)
  - Estimate pitches and polyphony in each frame
  - Many methods



- Note-level (note tracking)
  - Estimate pitch, onset, offset of notes
  - Fewer methods



- Stream-level (multi-pitch streaming)
  - Stream pitches by sources
  - Very few methods



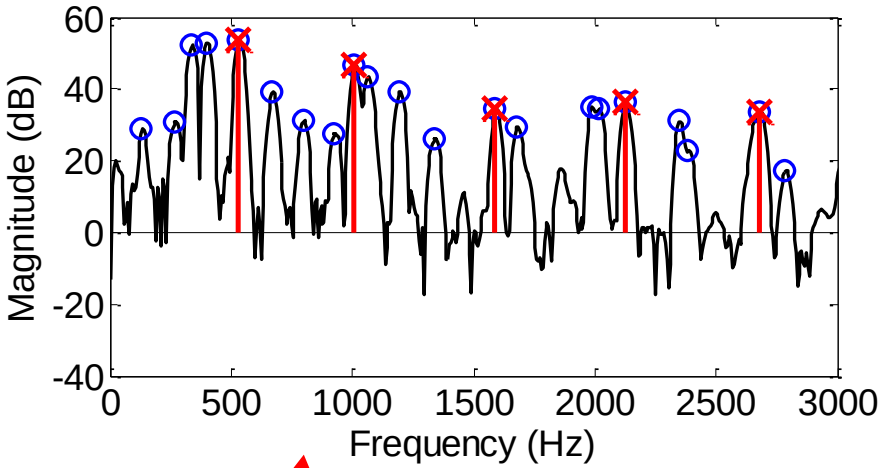
Multi-pitch Streaming (Timbre Tracking)

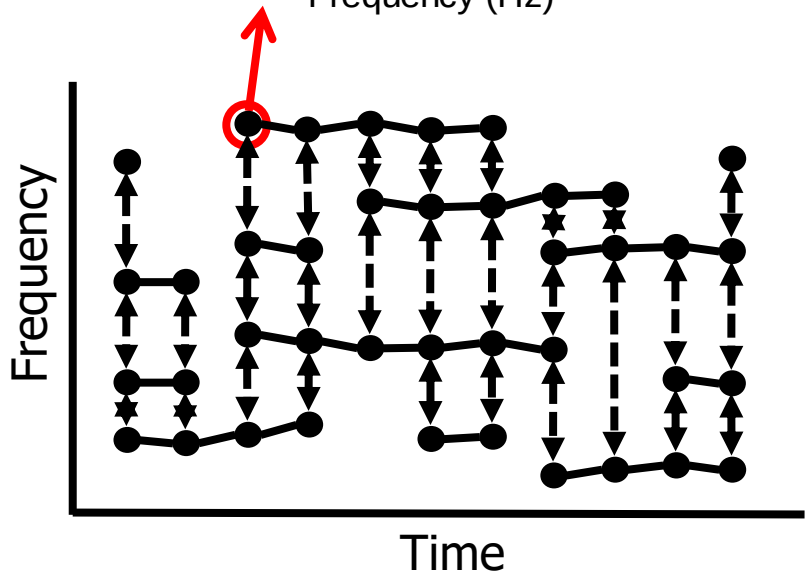
- Supervised
  - Train timbre models of sound sources
  - Apply timbre models during pitch estimation: [Cont et al., 2007; Bay et al., 2012; Benetos et al., 2013]
  - Classify estimated pitches/notes: [Wu et al. 2011]
- Supervised with timbre adaptation
  - Adapt trained timbre models to sources in mixture: [Carabias-Orti et al., 2011; Grindlay & Ellis, 2011]
- Unsupervised
  - Cluster pitch estimates according to timbre: [Duan et al., 2009, 2014; Mysore & Smaragdis, 2009; Arora & Behera, 2015]

Timbre Tracking – Unsupervised (1)

[Duan et al., 2009, 2014]

- Constrained clustering
  - Objective: maximize timbre consistency within clusters
  - Constraints based on pitch locations: must-links and cannot-links
- Timbre representation: harmonic structure feature
- Iterative algorithm: update clustering to monotonically decrease objective function and satisfy more constraints

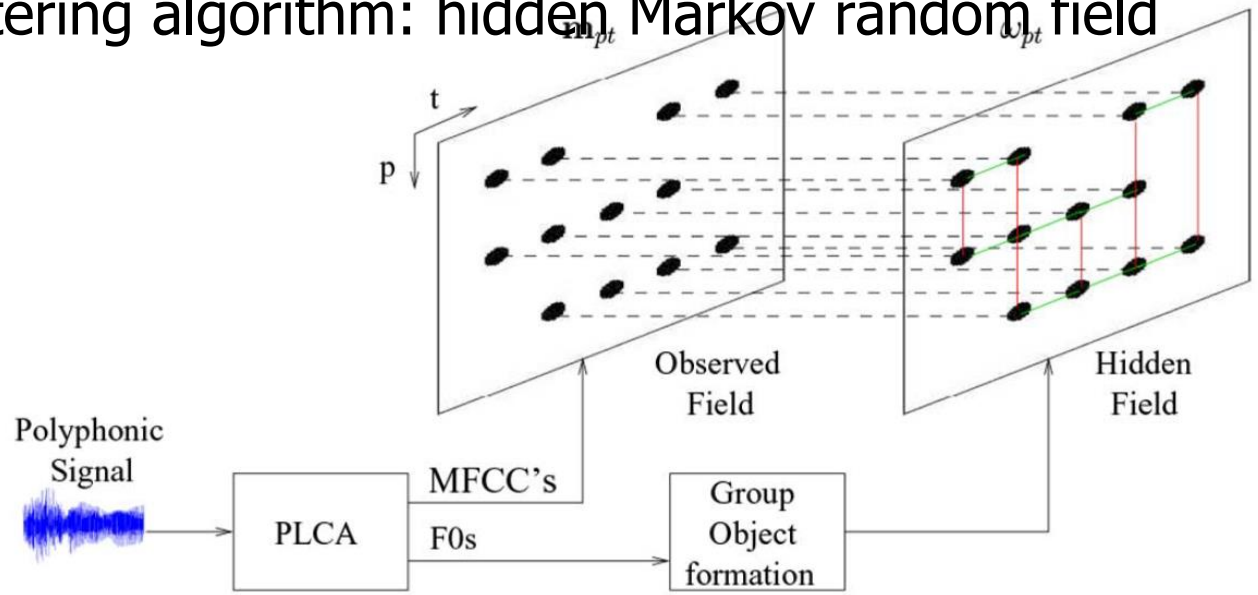




Timbre Tracking – Unsupervised (2)

[Arora & Behera, 2015]

- Constrained clustering
  - Objective: maximize timbre consistency within clusters
  - Constraints based on pitch locations: grouping constraints (i.e., pitch continuity) and simultaneity constraints (i.e., simultaneous pitches)
- Timbre representation: MFCC
- Clustering algorithm: hidden Markov random field

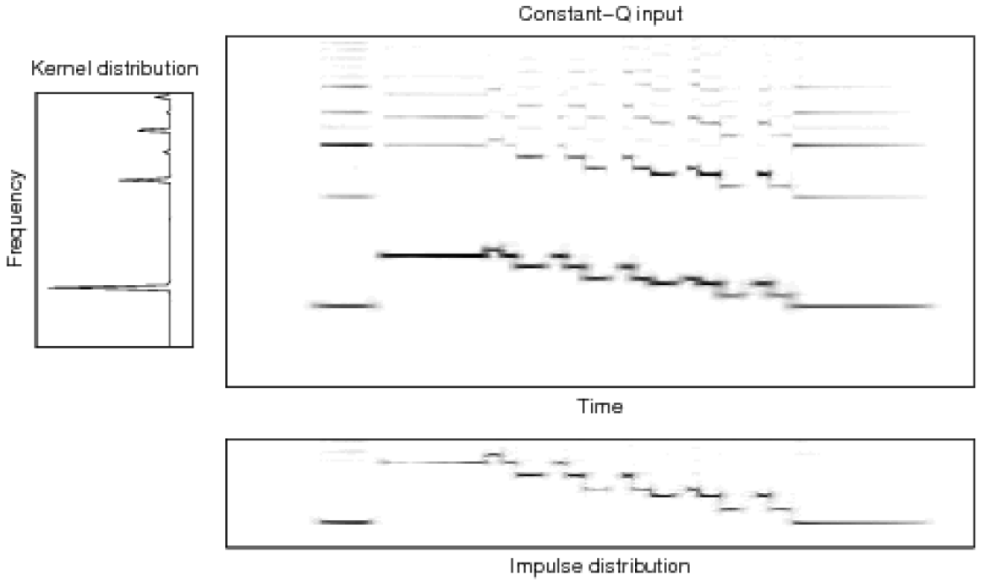


Timbre Tracking – Unsupervised (3)

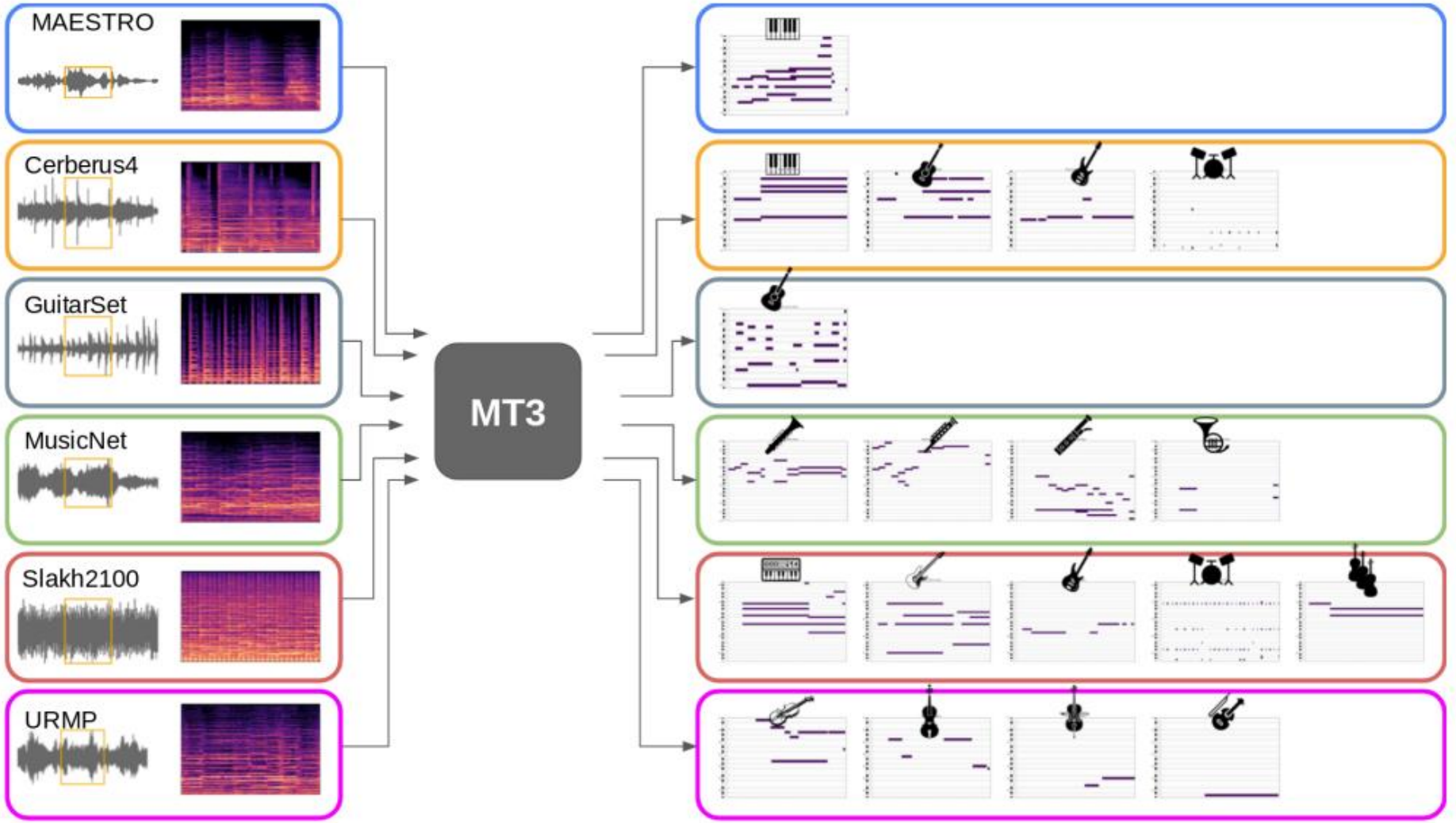
[Mysore & Smaragdis, 2009] for relative pitch tracking

- Shift-invariant PLCA on constant-Q spectrogram
  - Assumption: instrument spectrum shape invariant to pitch
  - Constraints: 1) note activation over frequency shift is unimodal; 2) note activation over time is smooth
- Can be viewed as a pitch clustering algorithm

- Pros: pitch estimation and timbre tracking are performed at the same time
- Cons: does not recognize the absolute pitch



MT3: Multi-Task Multitrack Music Transcription

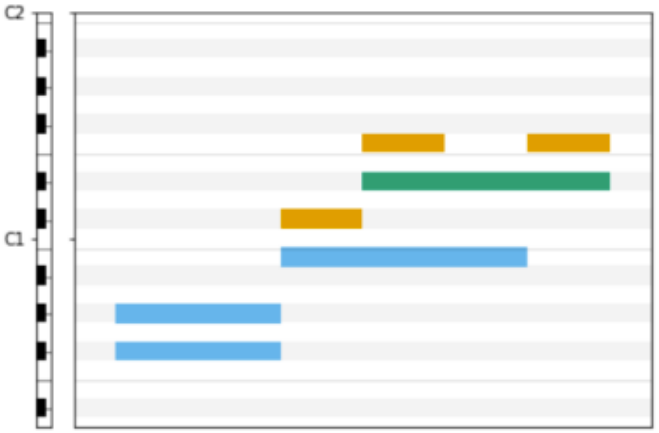


Josh Gardner, Ian Simon, Ethan Manilow, Curtis Hawthorne, Jesse Engel, "MT3: Multi-task multitrack music transcription," in *Proc. ICLR*, 2022.

Transformer Training

- Model input: log-mel spectrogram
- Model output: MIDI-like tokens

Piano Roll



Tokenization

Detokenization

MIDI-Like Target/Output Tokens

```
<time 0.27> <prog 26> <on>  <note 22>
                            <note 24>
<time 0.63>           <off> <note 22>
                            <note 24>
            <prog 0>  <on>  <note 29>
            <prog 26>       <note 27>
<time 1.55> <prog 32>       <note 31>
<time 1.60> <prog 0>  <off> <note 29>
                      <on>  <note 33>
<time 1.75>           <off> <note 33>
<time 1.90> <prog 26>       <note 27>
<time 2.03> <prog 32>       <note 31>
```

Dataset	Hrs. Audio	Num. Songs	Num. Instr.	Instr. Per Song	Align	Low-Resource	Synthetic	Drums
Slakh2100	969	1405	35	4–48	Good		✓	✓
Cerberus4	543	1327	4	4	Good		✓	✓
MAESTROv3	199	1276	1	1	Good			
MusicNet	34	330	11	1–8	Poor	✓		
GuitarSet	3	360	1	1	Good	✓		
URMP	1	44	14	2–5	Fair	✓		

Result Comparisons

Model	MAESTRO	Cerberus4	GuitarSet	MusicNet	Slakh2100	URMP
			Frame F1			
Hawthorne et al. (2021)	0.66	–	–	–	–	–
Manilow et al. (2020)	–	0.63	0.54	–	–	–
Cheuk et al. (2021)	–	–	–	0.48	–	–
Melodyne	0.41	0.39	0.62	0.13	0.47	0.30
MT3 (single dataset)	**0.88**	0.85	0.82	0.60	0.78	0.49
MT3 (mixture)	0.86	**0.87**	**0.89**	**0.68**	**0.79**	**0.83**
			Onset F1			
Hawthorne et al. (2021)	0.96	–	–	–	–	–
Manilow et al. (2020)	–	0.67	0.16	–	–	–
Cheuk et al. (2021)	–	–	–	0.29	–	–
Melodyne	0.52	0.24	0.28	0.04	0.30	0.09
MT3 (single dataset)	**0.96**	0.89	0.83	0.39	0.76	0.40
MT3 (mixture)	0.95	**0.92**	**0.90**	**0.50**	**0.76**	**0.77**

Summary

- Basic Concepts of Pitch
  - Pitch perception
  - Pitch and music
- Single Pitch Detection
  - Time domain
  - Spectral domain
  - Cepstral domain
  - Machine learning methods
- Multi-Pitch Analysis
  - Frame-level
  - Note-level
  - Stream-level


Thank You!