Topic 4

Single Pitch Detection

What is pitch?

- 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

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?

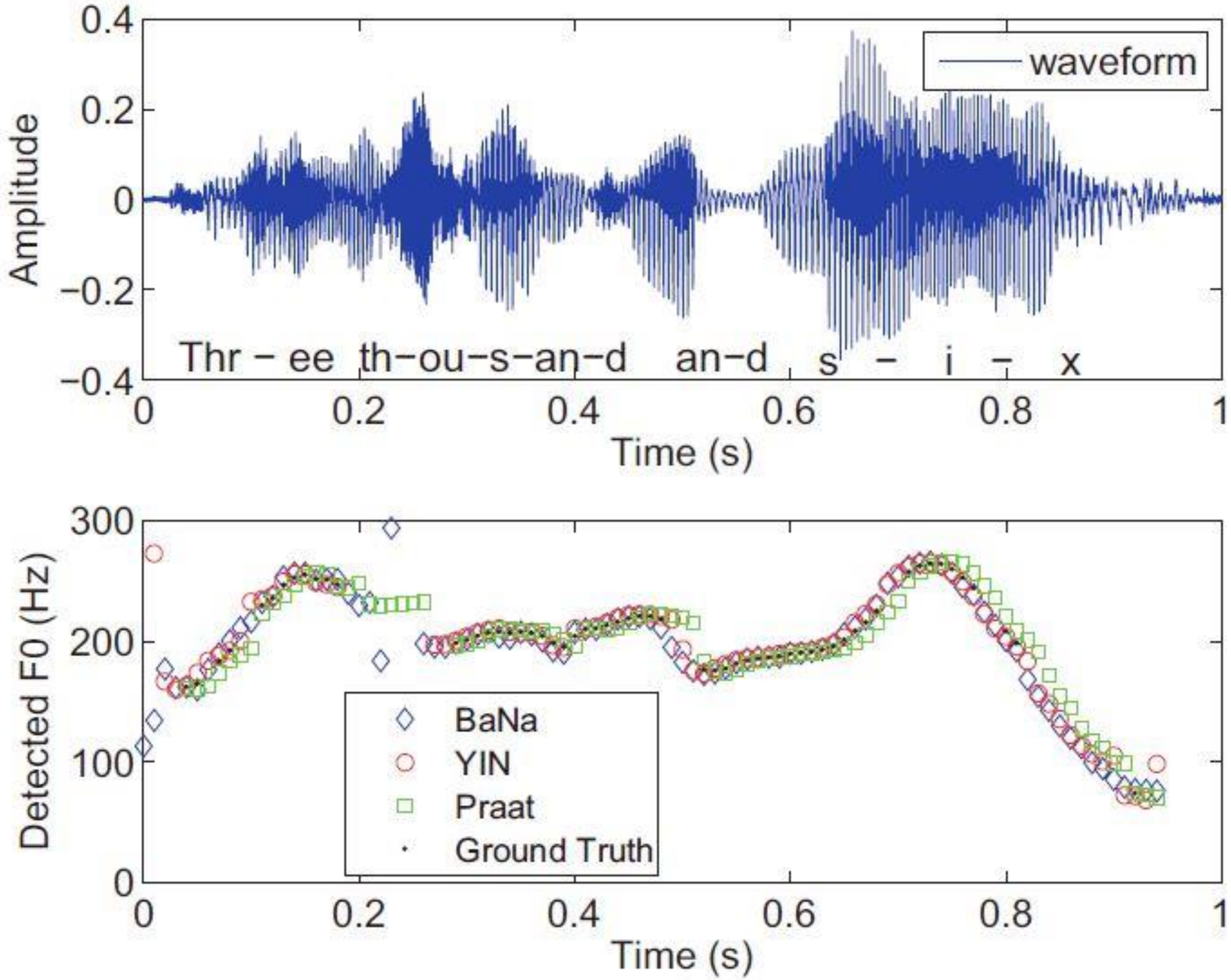
General Process of Pitch Detection

- 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

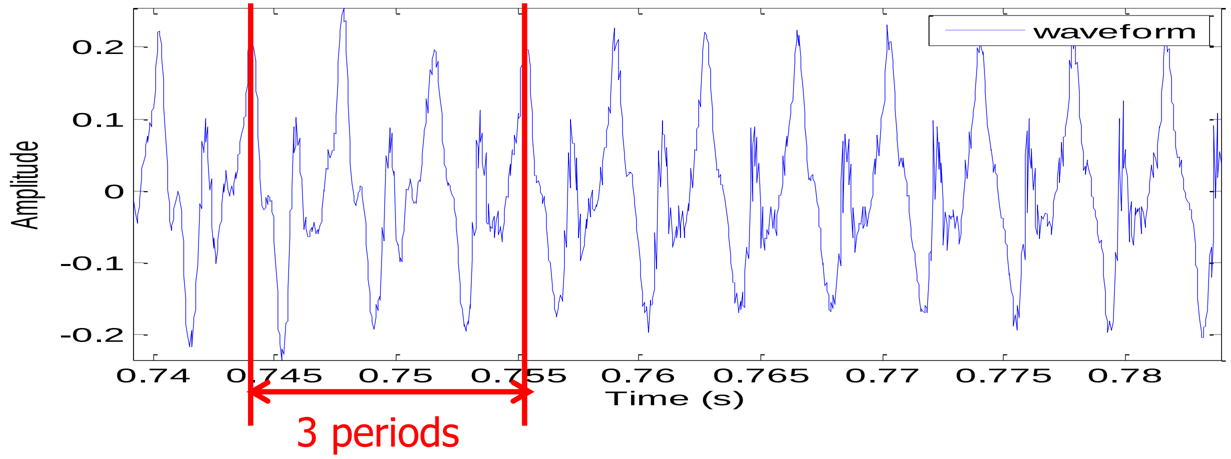
An Example



waveform
Amplitude
Thr − ee th−ou−s−an−d an−d s − i − x
Time (s)
Detected F0 (Hz)
BaNa
YIN
Praat
Ground Truth
Time (s)

How long should the frame be?

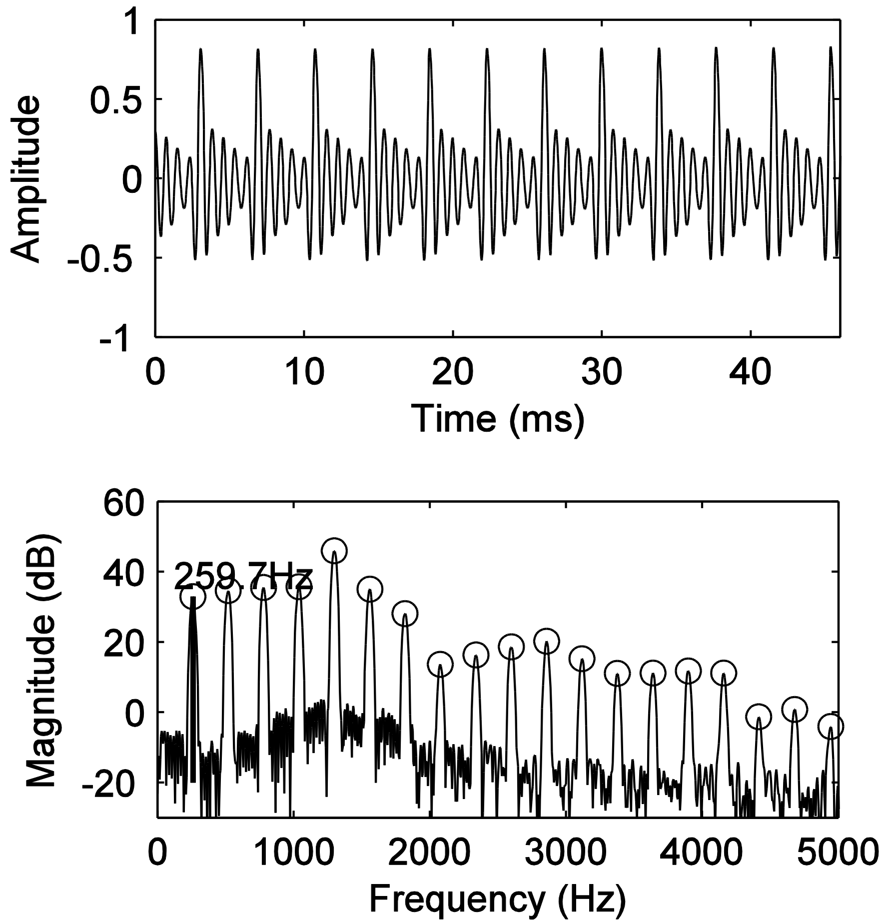
- Too long:
  - Contains multiple pitches (low time resolution)
- Too short
  - Can't obtain reliable detection (low freq resolution)
  - Should be at least about 2-3 periods of the signal



  - 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

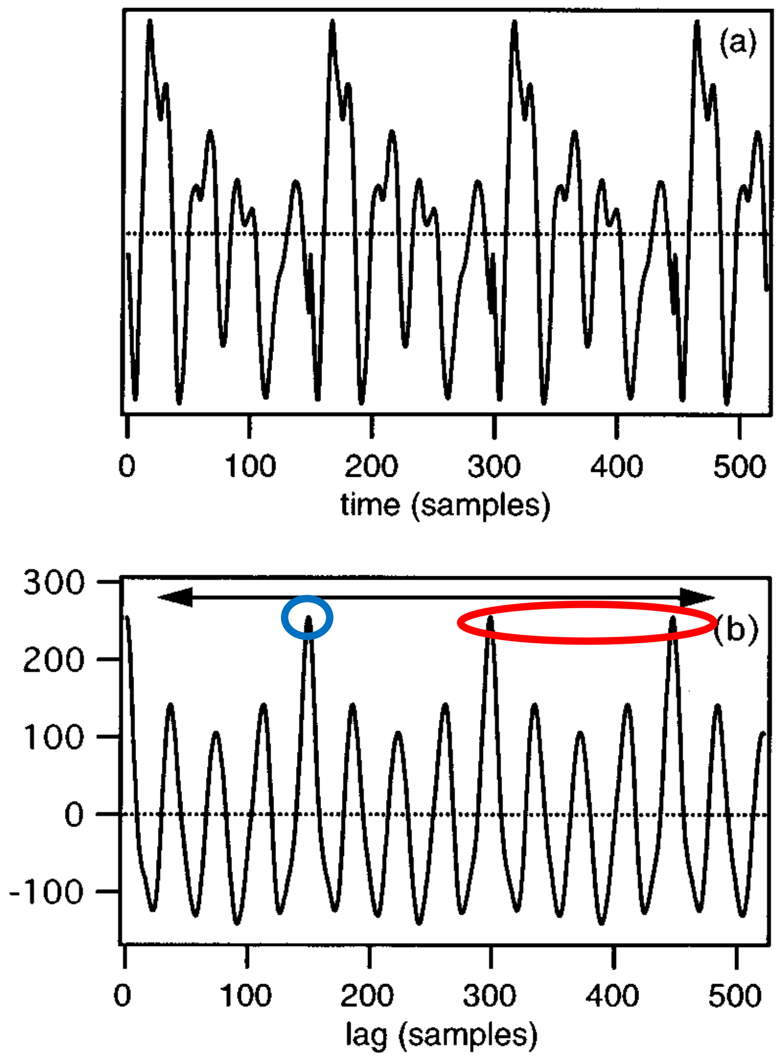
- Time domain signal is periodic
  - F0 = 1/period
- Time domain
  - Detect period
- Spectral peaks have harmonic relations
  - F0 is the greatest common divisor
- Frequency domain
  - Detect the divisor
- Spectral peaks are equally spaced
  - F0 is the frequency gap
- Cepstrum domain
  - Detect the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



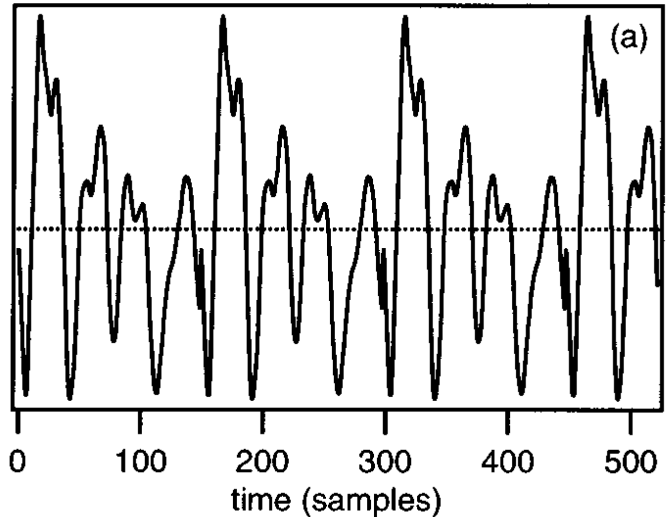
YIN – Step 2

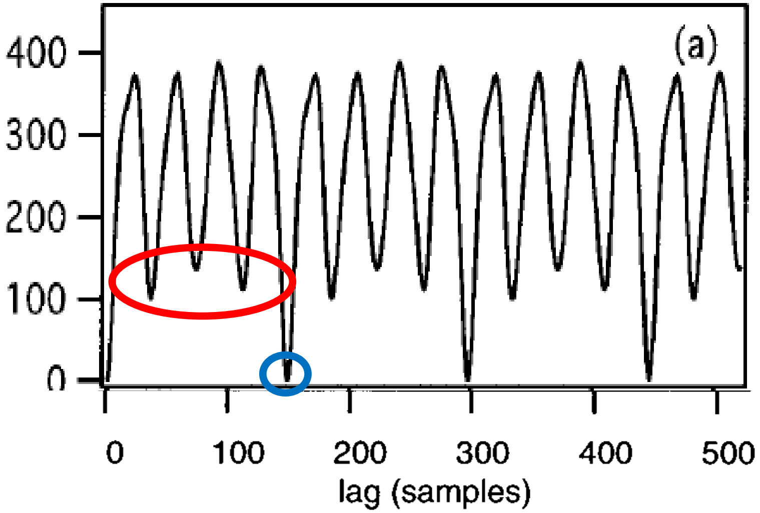
- Replace ACF with difference function

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

- 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

[de Cheveigne, 2002]

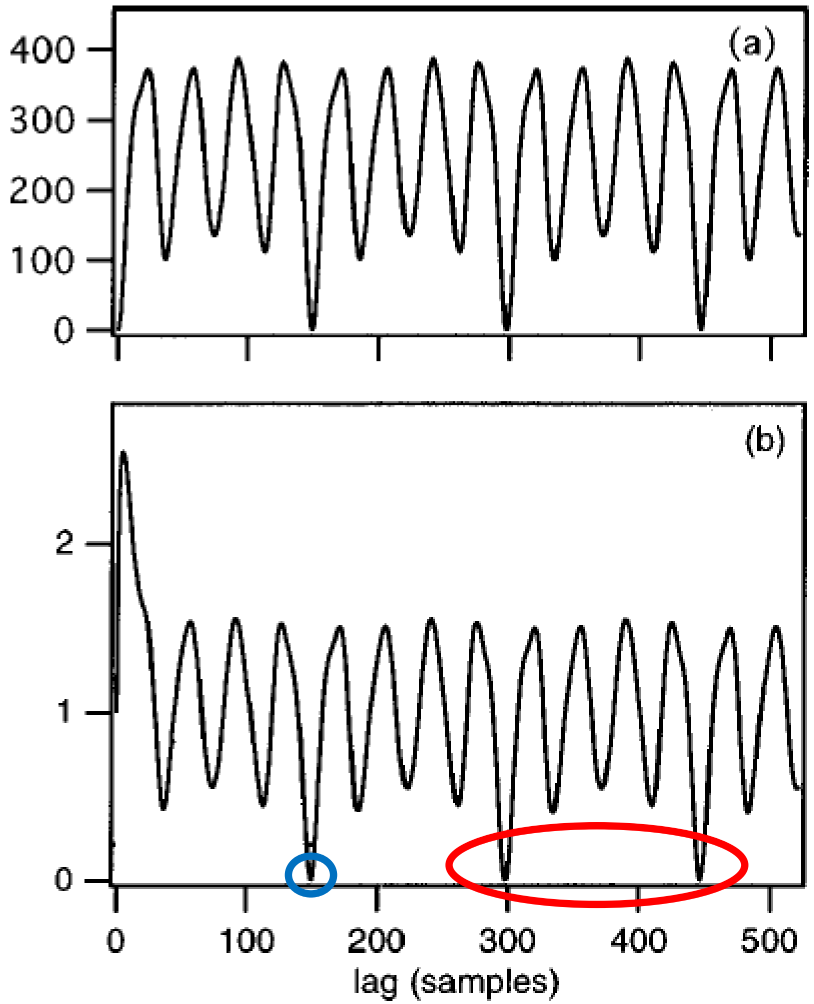




YIN – Step 3

- Cumulative mean normalized difference function

$$d'_t(\tau) = \begin{cases} 1, & \text{if } \tau = 0, \\ d_t(\tau) \bigg/ \left[(1/\tau) \sum_{j=1}^{\tau} d_t(j) \right] & \text{other} \end{cases}$$

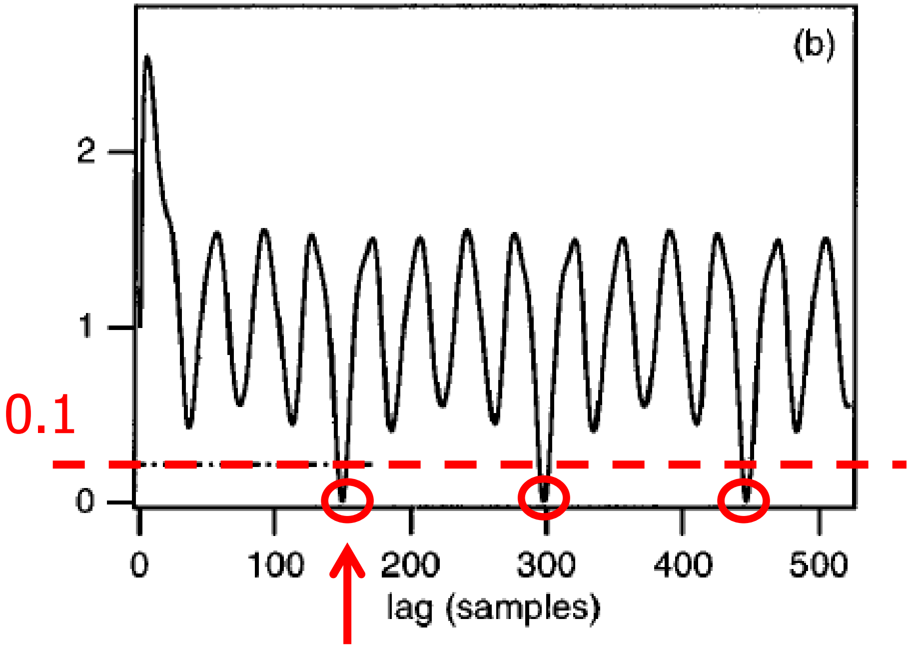


- Then take the deepest dip?

- Problem
  - May choose higher-order dips → lower octave error (or sub-harmonic error)

YIN – Step 4

- Absolute Threshold
  - Set threshold to say 0.1
  - Pick the first dip that exceeds the threshold



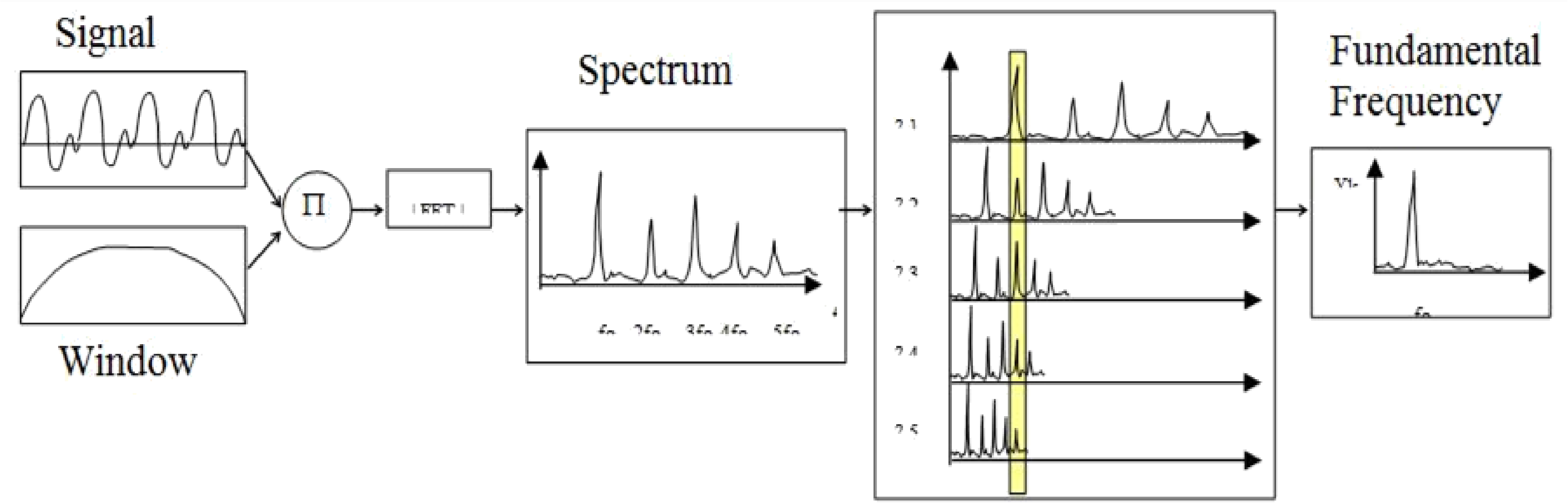
YIN – Step 5 & 6

- Step 5: parabolic interpolation to find the exact dip location
  - The dip location in the discrete world may deviate from the exact dip location
- Step 6: use the best local estimate
  - Some analysis points may be better than others (result in smaller d’)
  - Use the pitch estimate from the best analysis point within the frame

Frequency Domain Approach

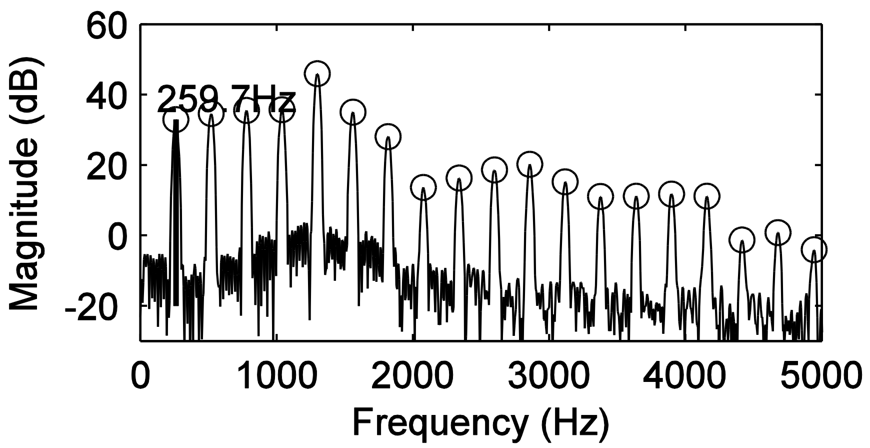
- Idea: for each F0 candidate, calculate the support (e.g., spectral energy) it receives from its harmonic positions.
- Harmonic Product Spectrum (HPS)

[Schroeder, 1968; Noll, 1970]



Cepstral Domain Approach

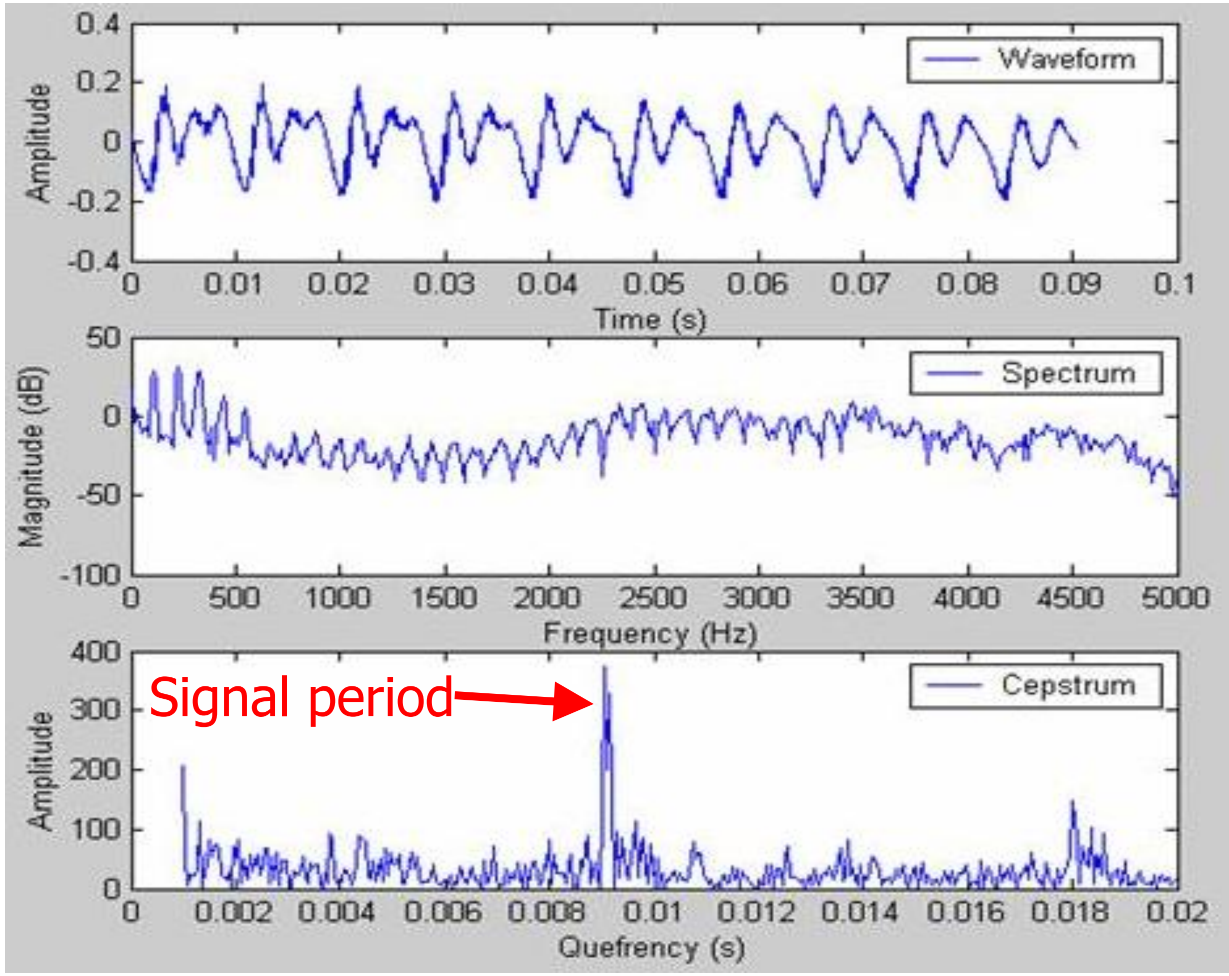
- Idea: find the frequency gap between adjacent spectral peaks
  - The log-amplitude spectrum looks pretty periodic
  - The gap can be viewed as the period of the spectrum
  - How to find the period then?
  - Cepstrum's idea: Fourier transform!



Cepstrum

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

Spectrum - Cepstrum
Frequency - Quefrency
Filtering - Liftering



Pitched or Non-pitched?

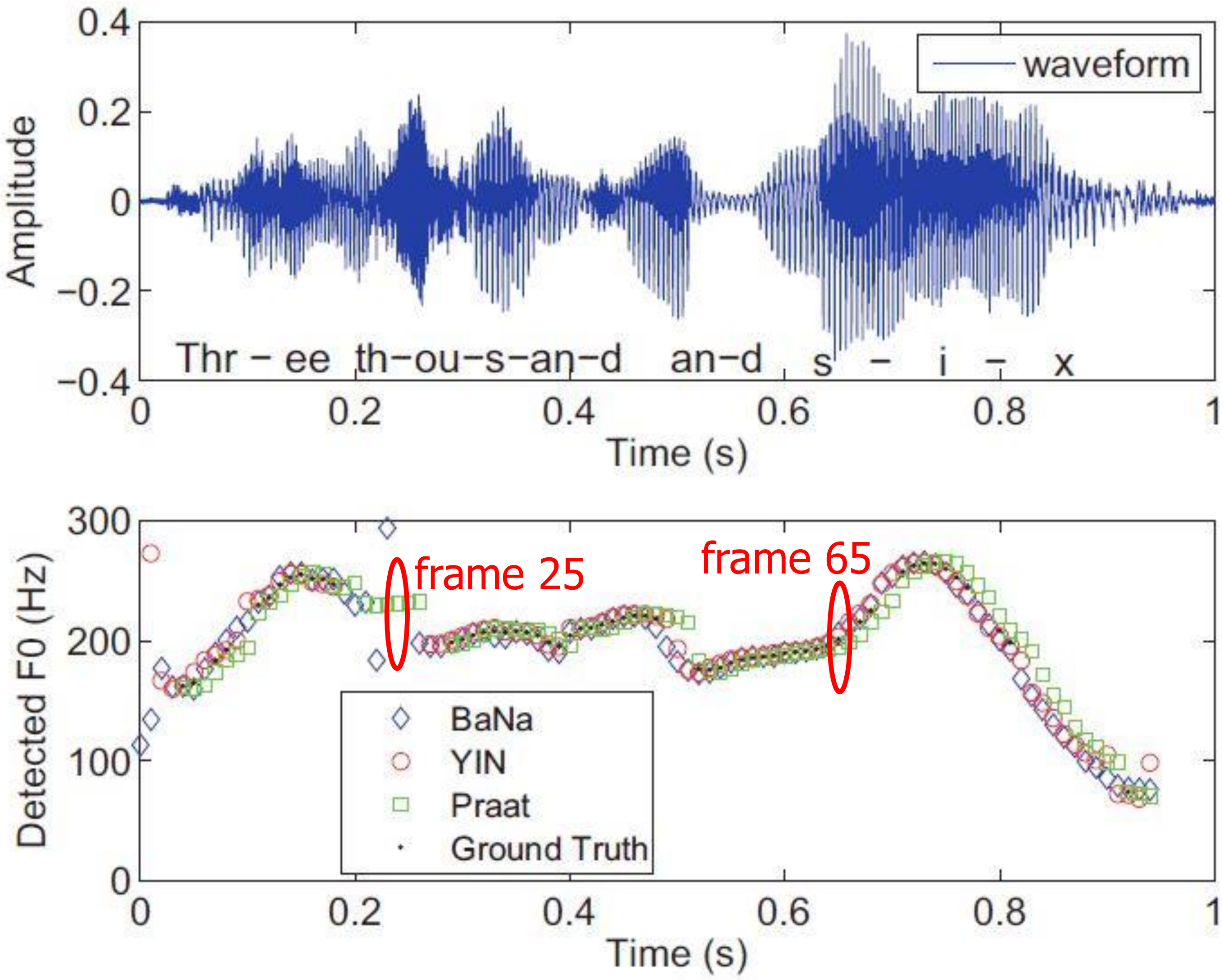
- Some frames may be silent or inharmonic, so they may not contain a pitch at all
  - Silence can be detected by RMS value
  - How about inharmonic frames?

- 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?

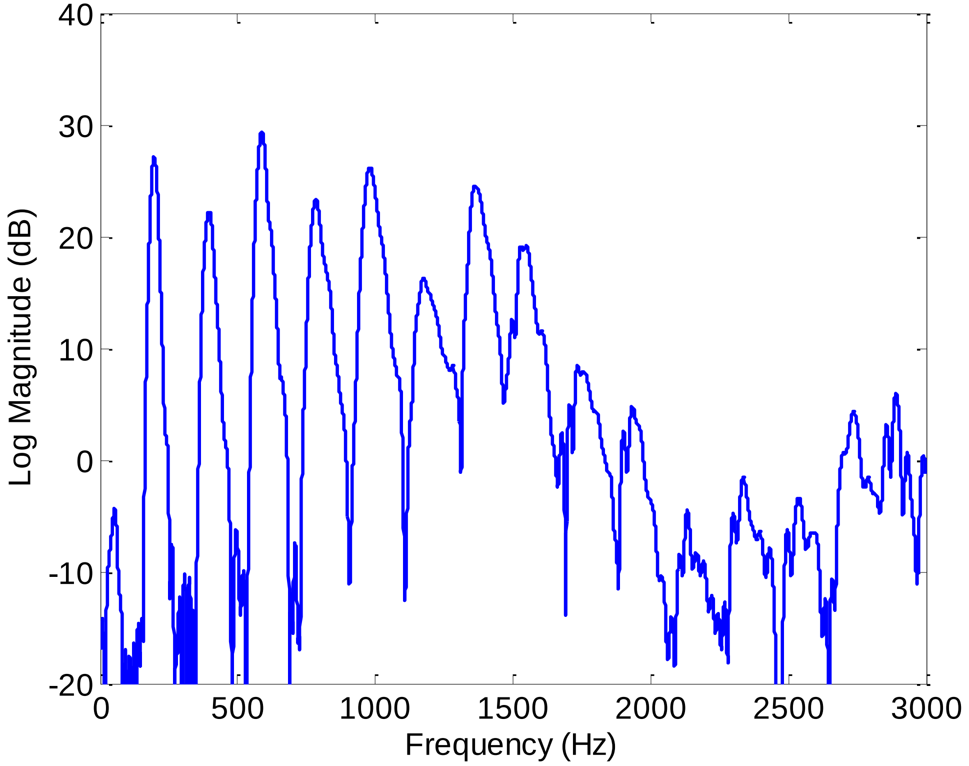
- Choose some recordings (speech, music)
- Get ground-truth
  - Listen to the signal and inspect the spectrum to manually annotate (time consuming!)
  - Automatic annotation using simultaneously recorded laryngograph signals for speech (not quite reliable!)
- Pitched/non-pitched classification error
- 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



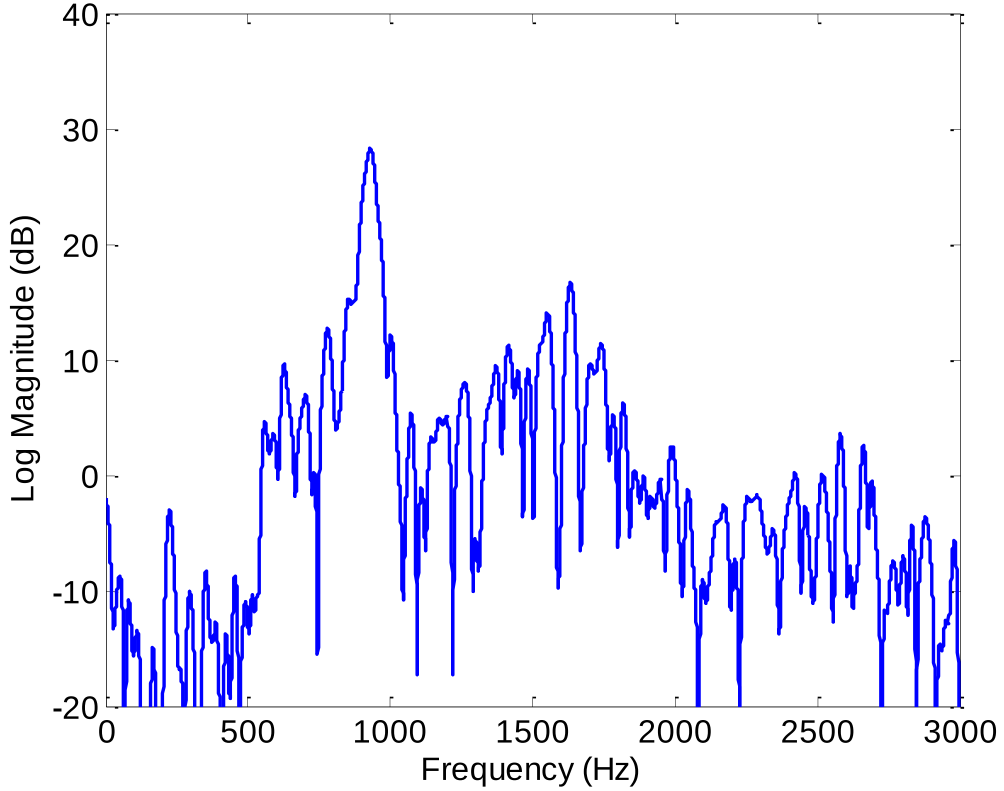
Frame 65 – Pitched (Voiced)

- Has clear harmonic patterns
- Different methods give close results, and consistent to the ground-truth 196 Hz



Frame 25 – Non-pitched (Unvoiced)

- No clear harmonic patterns
- Different methods give inconsistent results



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?

Summary

- Pitch detection is important for many tasks
  - Time domain: find the period of waveform
  - Frequency domain: find the common divisor of peak frequencies
  - Cepstral domain: find the frequency gap between spectral peaks

- Pitch detection research is quite mature in noiseless conditions

- Pitch detection in noisy environments (also called robust pitch detection, noise-resilient pitch detection) is an active research topic
  - BaNa [Yang et al., 2014]; PEFAC [Gonzales & Brookes, 2014]; Crepe [Kim et al., 2018]; SPICE [Gfeller et al., 2019]

References

- Childers, D. G., Skinner, D.P., and Kemerait, R.C. (1977). The cepstrum: A guide to processing. In *Proc. IEEE*.
- de Cheveigne, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. JASA.
- Noll, A. M. (1970). Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum and a maximum likelihood estimate. In *Proc. SCPC*.
- Rabiner, L. R., Cheng, M. J., Osenberg, A. E., & McGonegal, C. A. (1976). A comparative performance study of several pitch detection algorithms. *TASSP*.
- Schroeder, M. R. (1968). Period histogram and product spectrum: New methods for fundamental frequency measurement. *JASA*.
- Yang, N., Ba, H., Demirkol, I., & Heinzelman, W. (2014). A noise resilient fundamental frequency detection algorithm for speech and music. *TASLP*.
- Gonzalez, S., & Brookes, M. (2014). PEFAC - a pitch estimation algorithm robust to high levels of noise. TASLP.
- Jong Wook Kim, Justin Salamon, Peter Li, Juan Pablo Bello. (2018). CREPE: A convolutional representation for pitch estimation. In *Proc. ICASSP.*
- Beat Gfeller, Christian Frank, Dominik Roblek, Matt Sharifi, Marco Tagliasacchi, and Mihajlo Velimirović. (2020). SPICE: Self-supervised pitch estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1118-1128, 2020.