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 language), statement/question, emotion, etc.

What scales are used?

What emotion?

mā má mǎ mà
mom numb horse scold

妈 麻 马 骂
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
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.
  - $F_0 = 1/\text{period}$

- Spectral peaks have harmonic relations.
  - $F_0$ is the greatest common divisor

- Spectral peaks are equally spaced.
  - $F_0$ is the frequency gap
Pitch Detection Methods

- Time domain signal is periodic.
  - \( F_0 = \frac{1}{\text{period}} \)

- Spectral peaks have harmonic relations.
  - \( F_0 \) is the greatest common divisor

- Spectral peaks are equally spaced.
  - \( F_0 \) 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 – Step 2

- Replace ACF with difference function

\[
d_i(\tau) = \sum_{j=1}^{W} (x_j - x_{j+\tau})^2
\]

- 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 – Step 3

- Cumulative mean normalized difference function

\[ d'_{t}(\tau) = \begin{cases} 
1, & \text{if } \tau = 0, \\
\frac{d_t(\tau)}{\left[ \frac{1}{\tau} \sum_{j=1}^{\tau} d_t(j) \right]} & \text{other}
\end{cases} \]

- Then take the deepest dip?

- Problem
  - May choose higher-order dips \(\rightarrow\) 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

\[
\text{power cepstrum} = |\mathcal{F}^{-1}\{\log|\mathcal{F}\{x(t)\}|^2\}|^2
\]
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.
- 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?

• Will pitch detection algorithms still work?

• Which domain is less sensitive to which kind of noise?

• How to improve pitch detection in noisy environments?

Violin + babble noise
Summary

• Pitch detection is important for many tasks
  – Time domain: find the period of waveform
  – Frequency domain: find the divisor of peaks
  – Cepstral domain: find the frequency gap between spectral peaks

• Pitch detection research is pretty 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];
References


• de Cheveigne, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. JASA.


