Topic 4

Single Pitch Detection

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



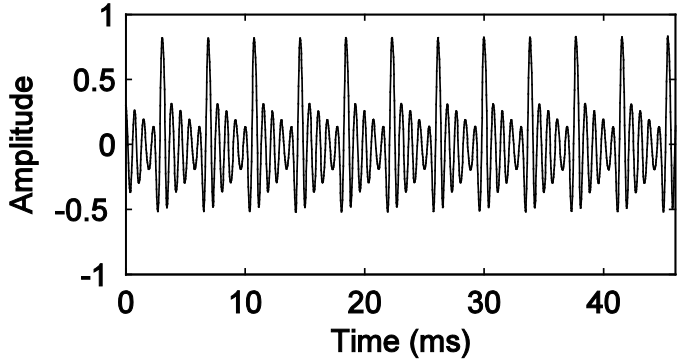
Harmonic Sound

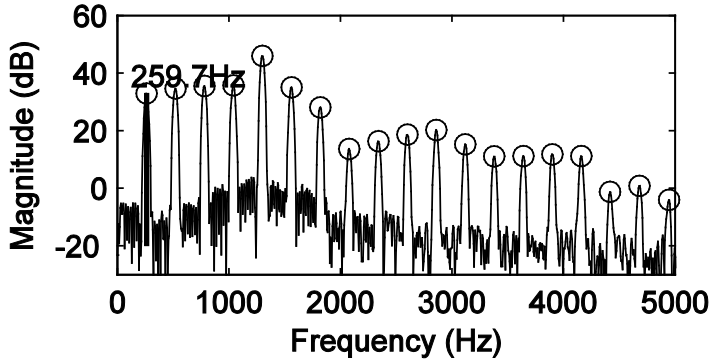
- A sound with strong sinusoidal components at integer multiples of a fundamental frequency. These components are called **harmonics** or **overtones**.

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

Classify Sounds by Harmonicity

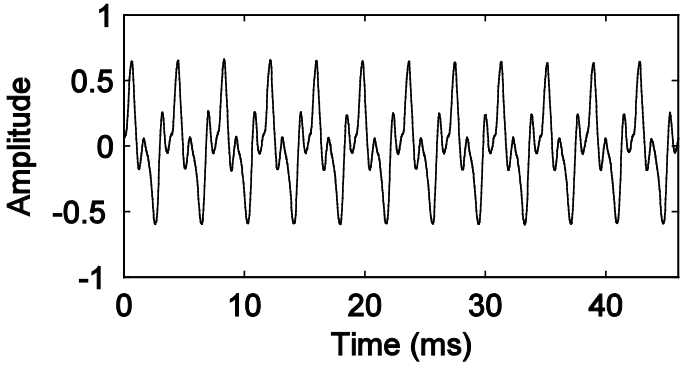
- Sine wave
- Strongly harmonic

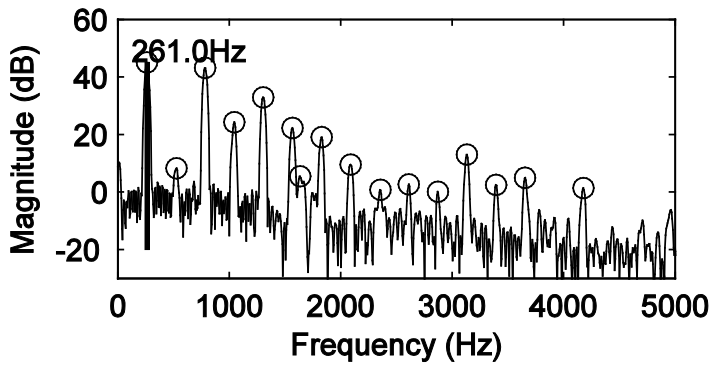




Oboe





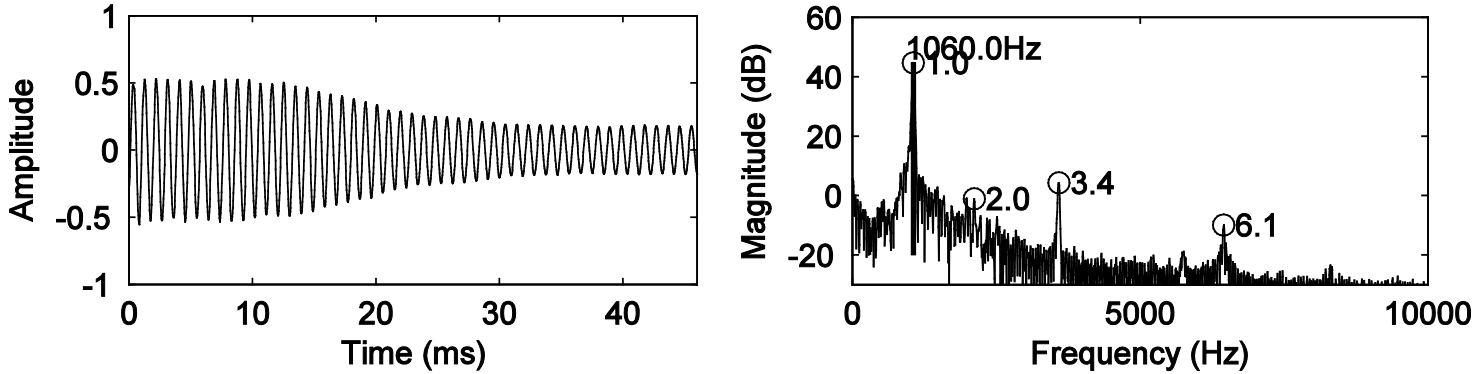


Clarinet

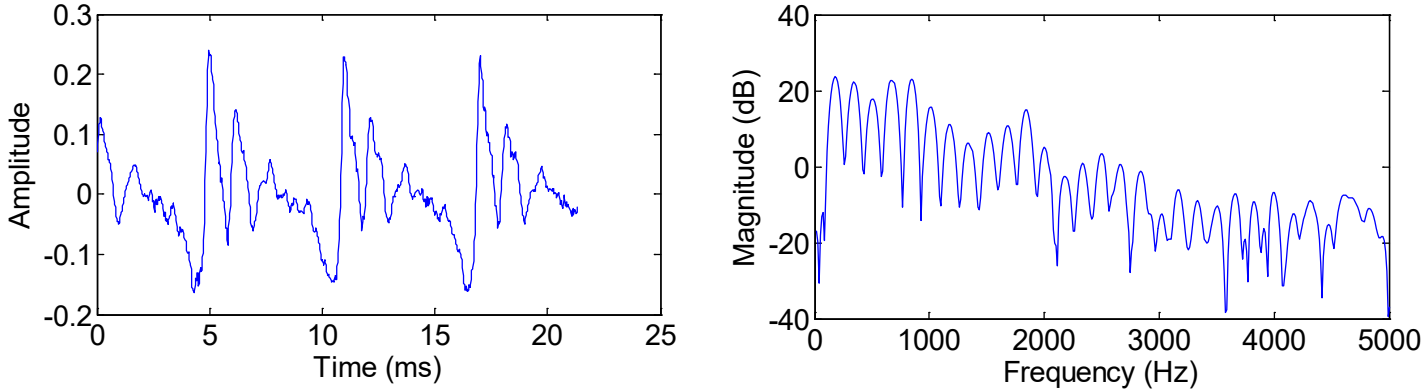


Classify Sounds by Harmonicity

- Somewhat harmonic (quasi-harmonic)



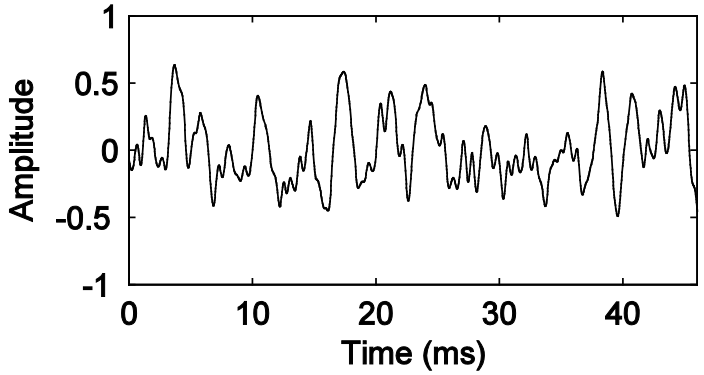
Marimba

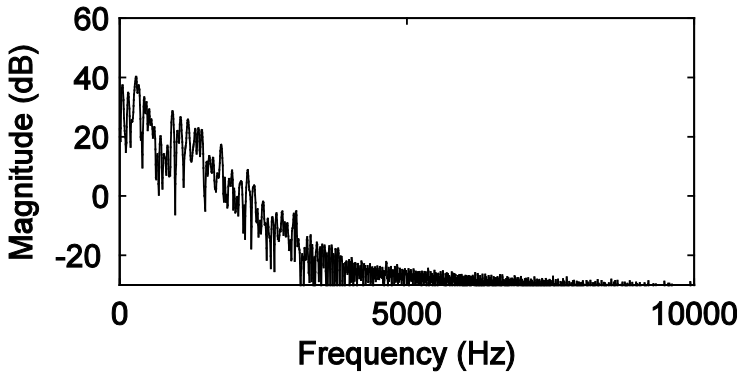


Human voice

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.)

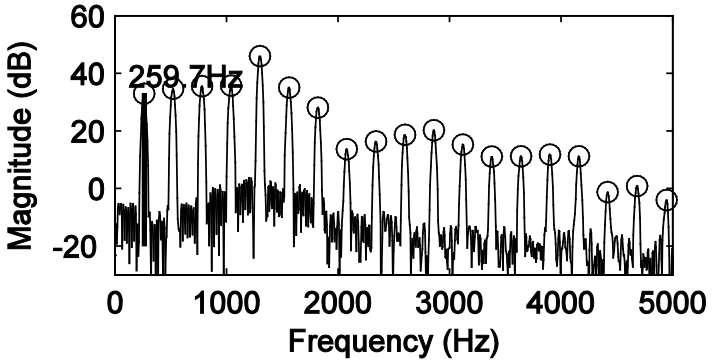
What determines pitch?

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

- Let’s listen and explore…

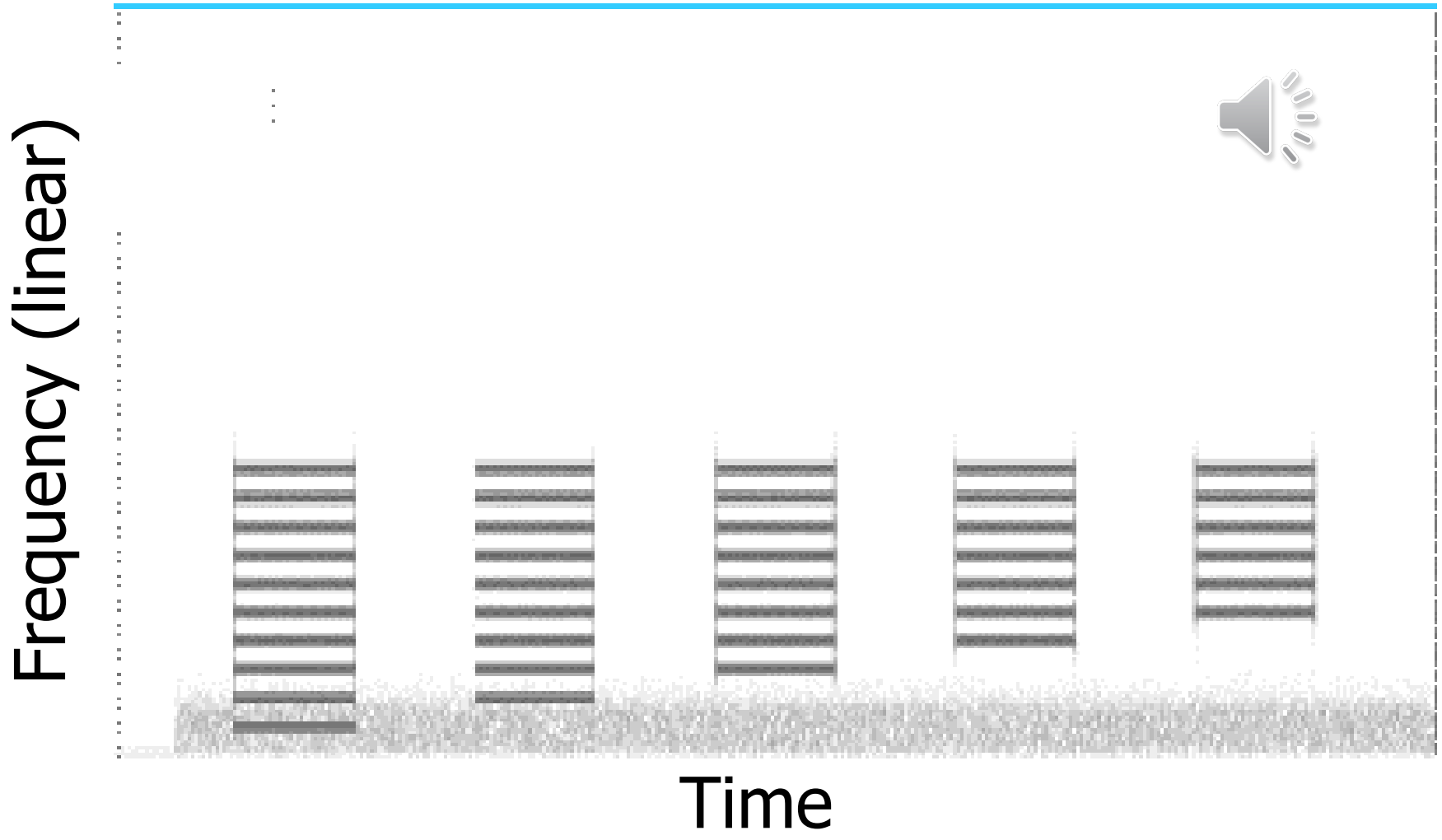
Hypothesis

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



Oboe

The Missing Fundamental



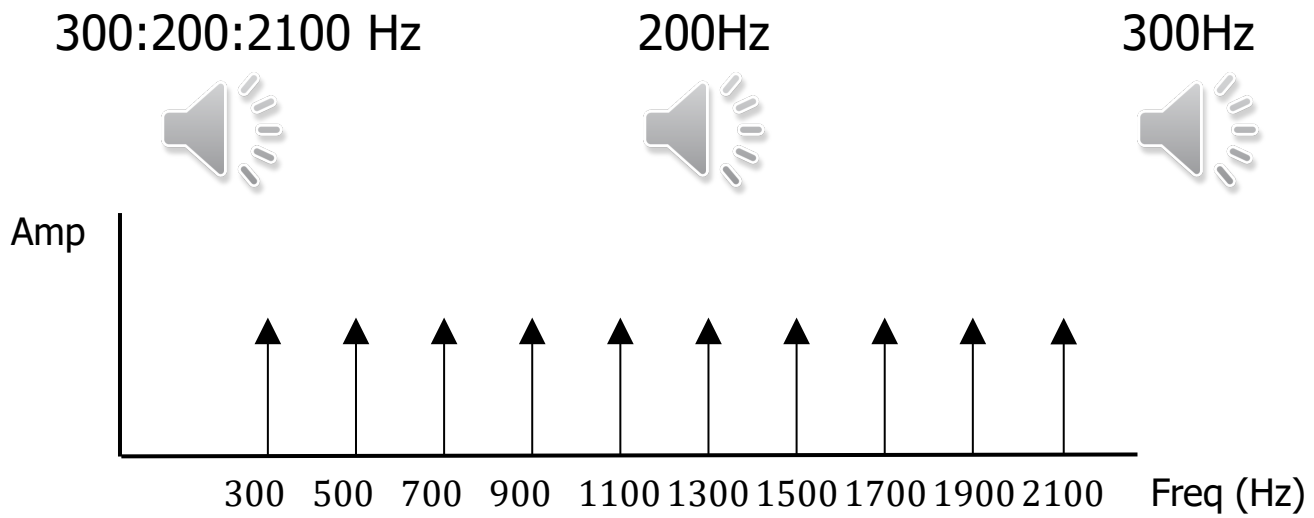
Hypothesis

- 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

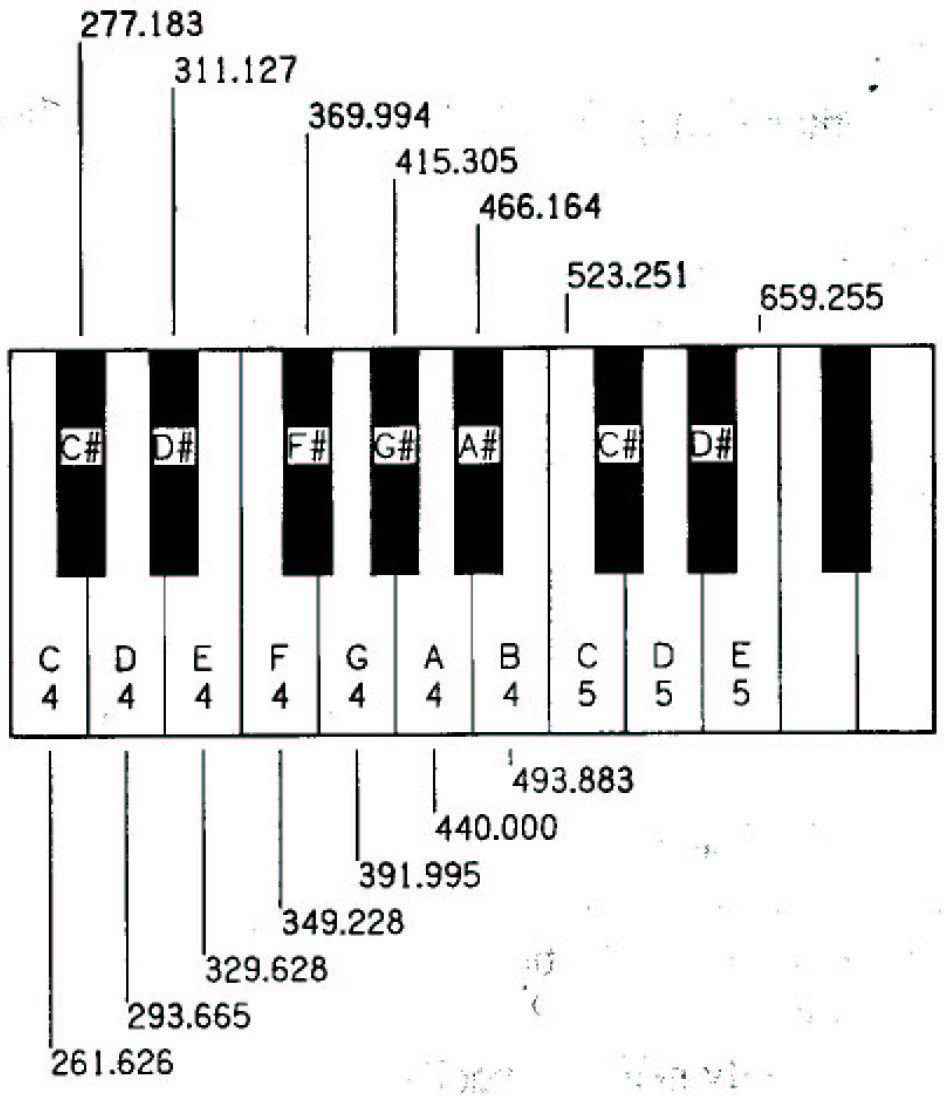
Measurement

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

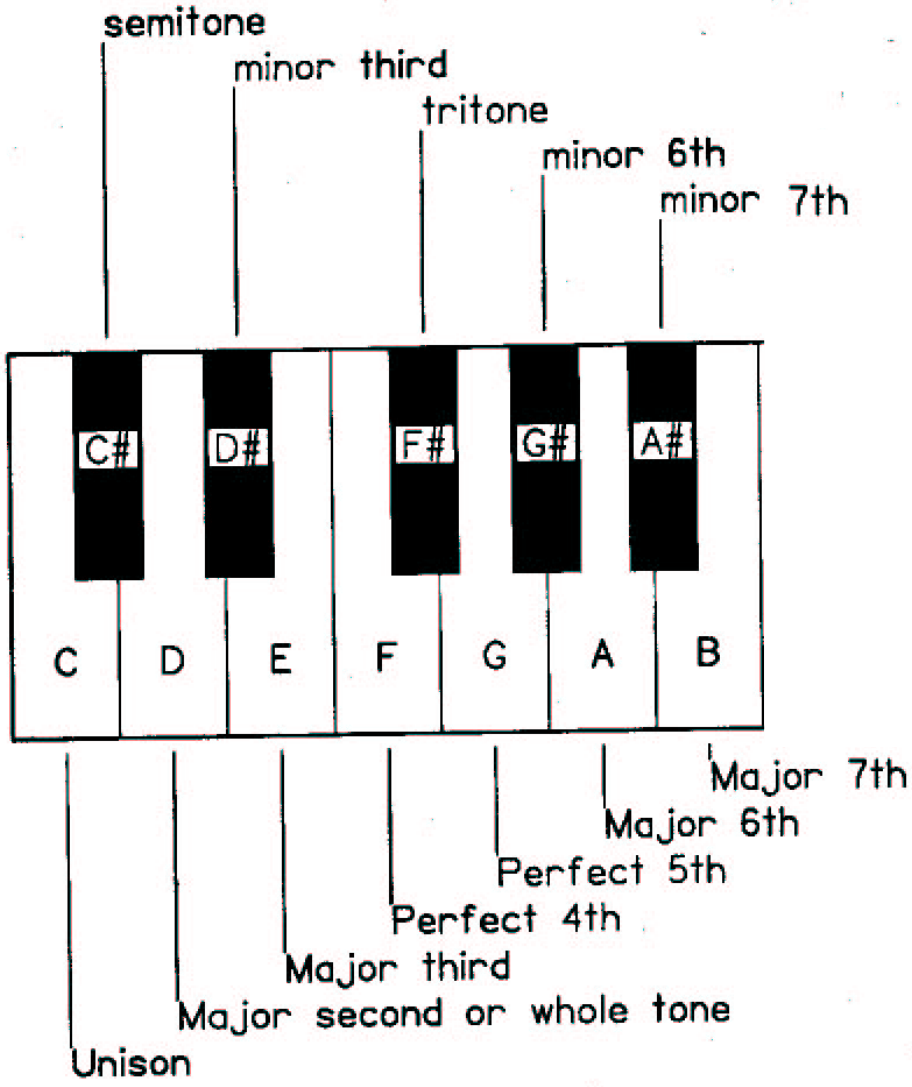
Number of cents

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

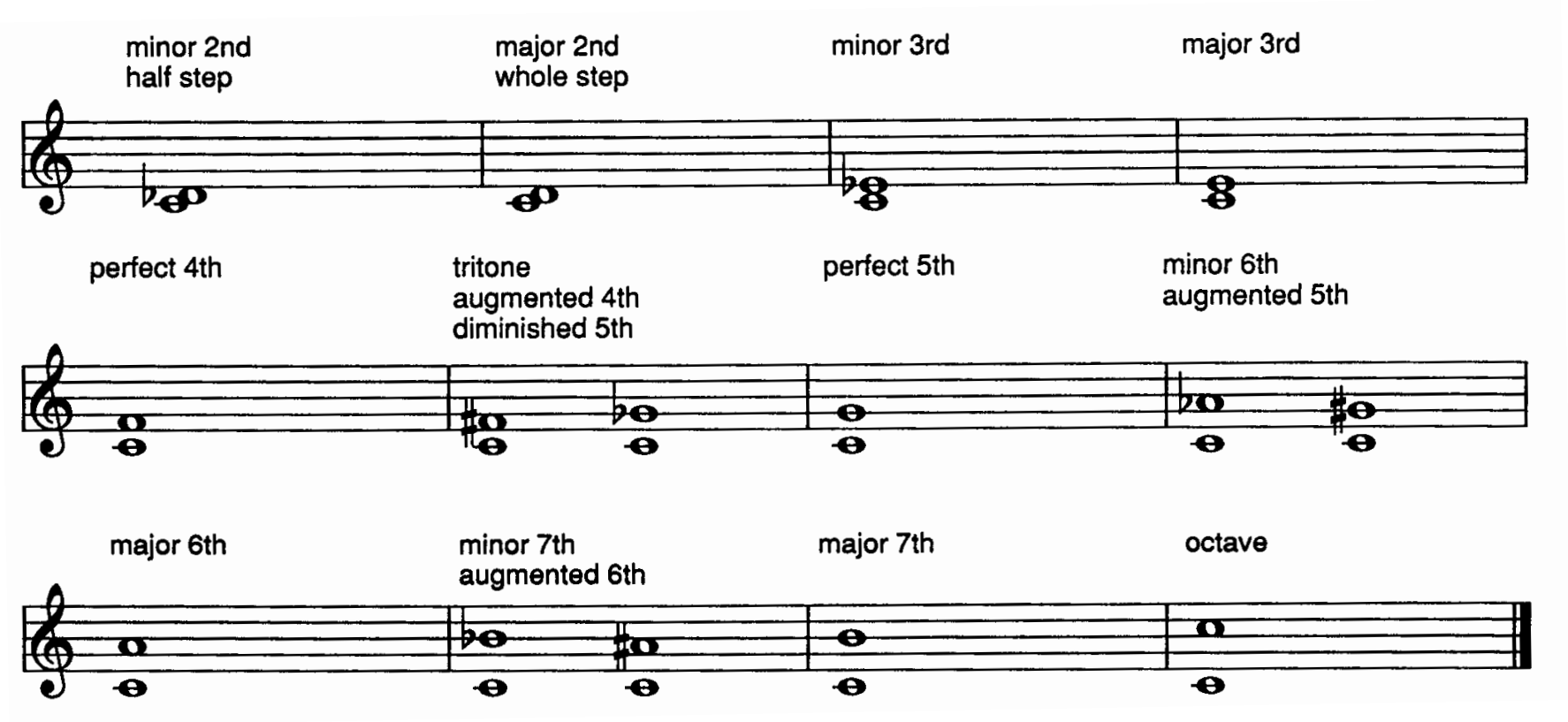
A=440 Equal Temperament Tuning



Musical Intervals (from C)

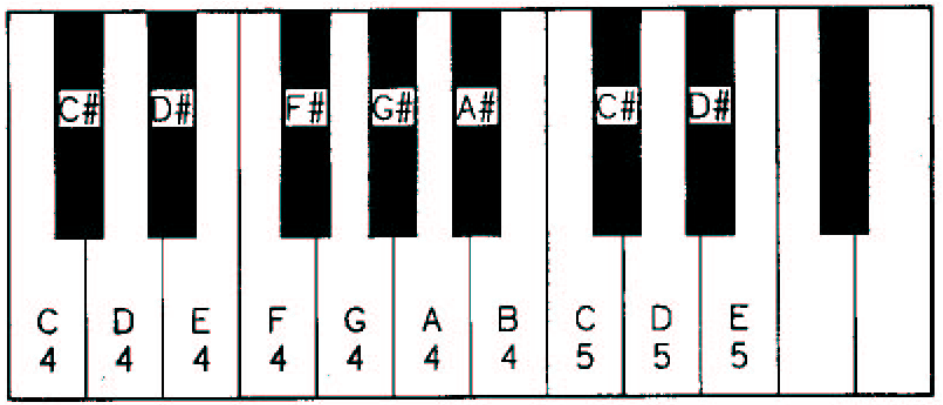


Interval Names



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, $2^{\frac{7}{12}} = 1.4983 \approx \frac{3}{2}$

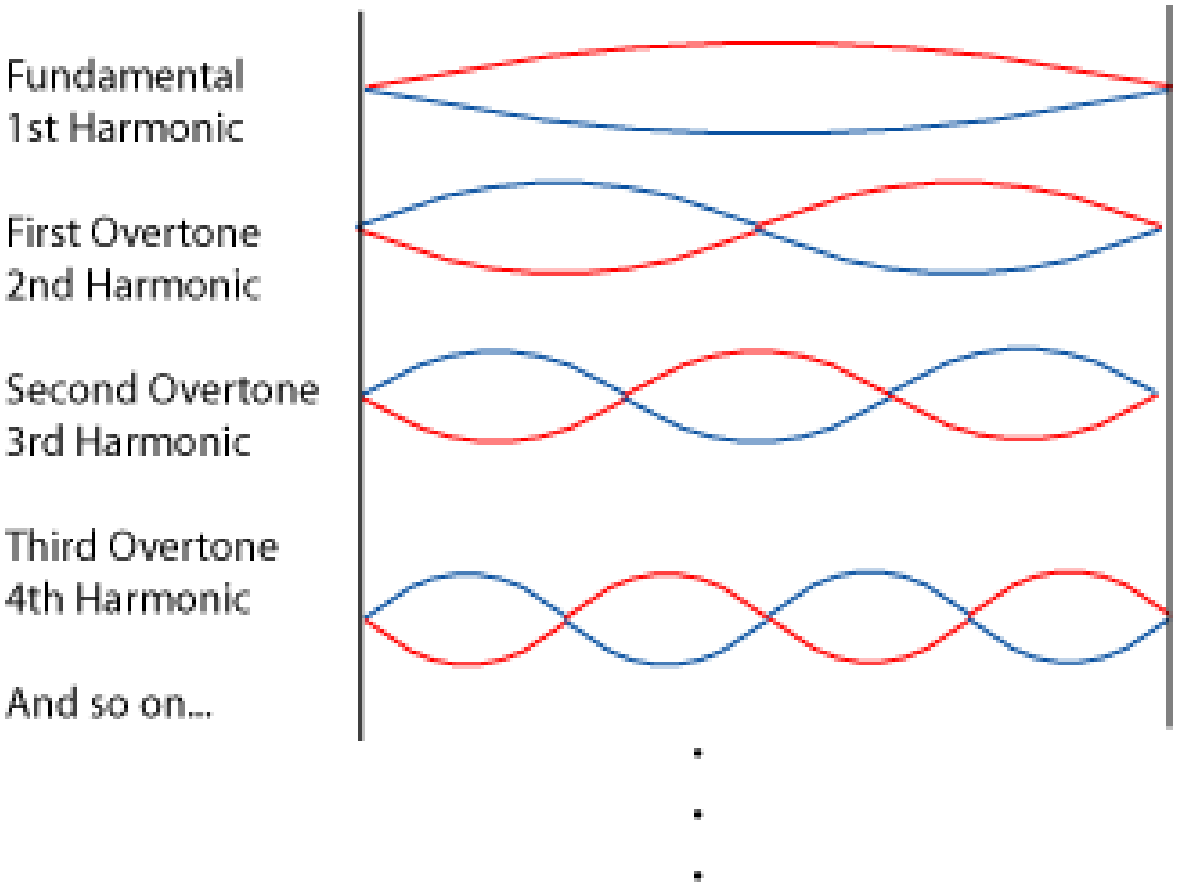
C → F: 5 half-steps, $2^{\frac{5}{12}} = 1.3348 \approx \frac{4}{3}$

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

Are these just coincidences?

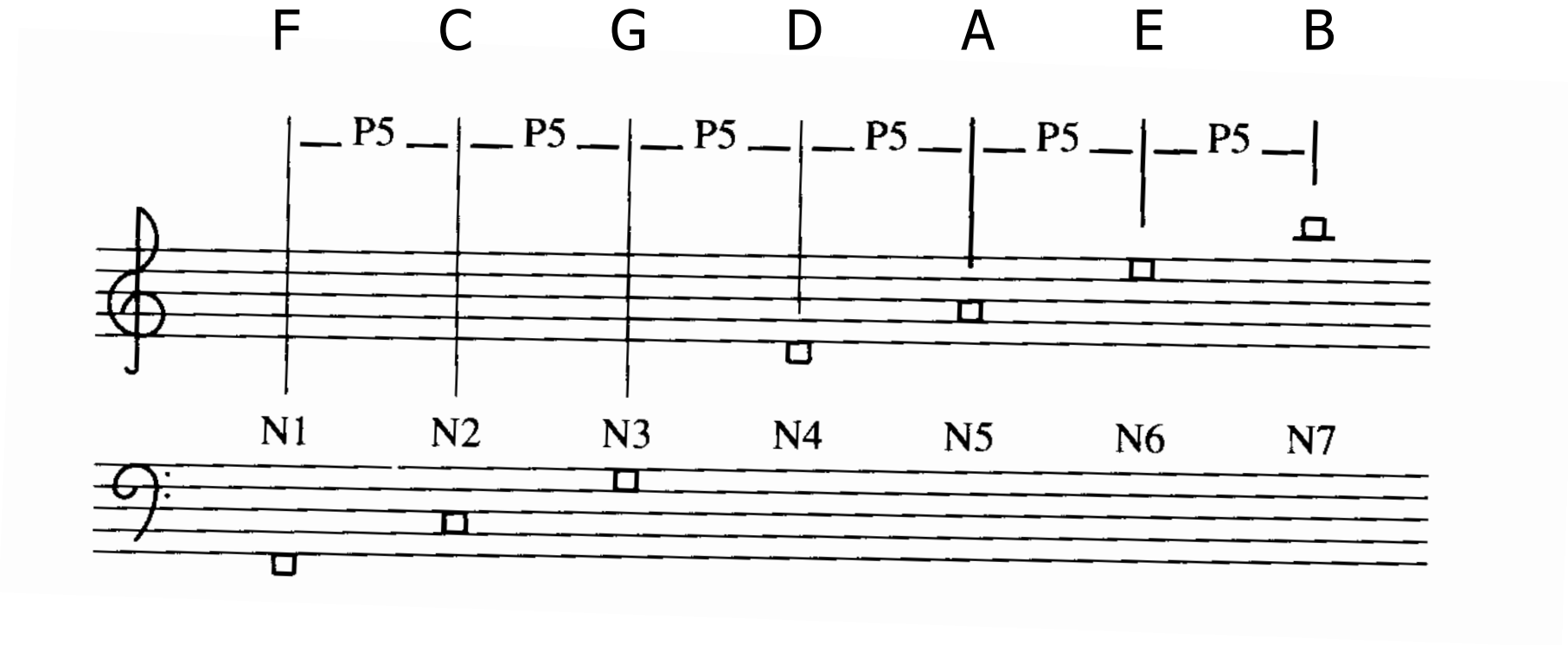
Related to Standing Waves

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

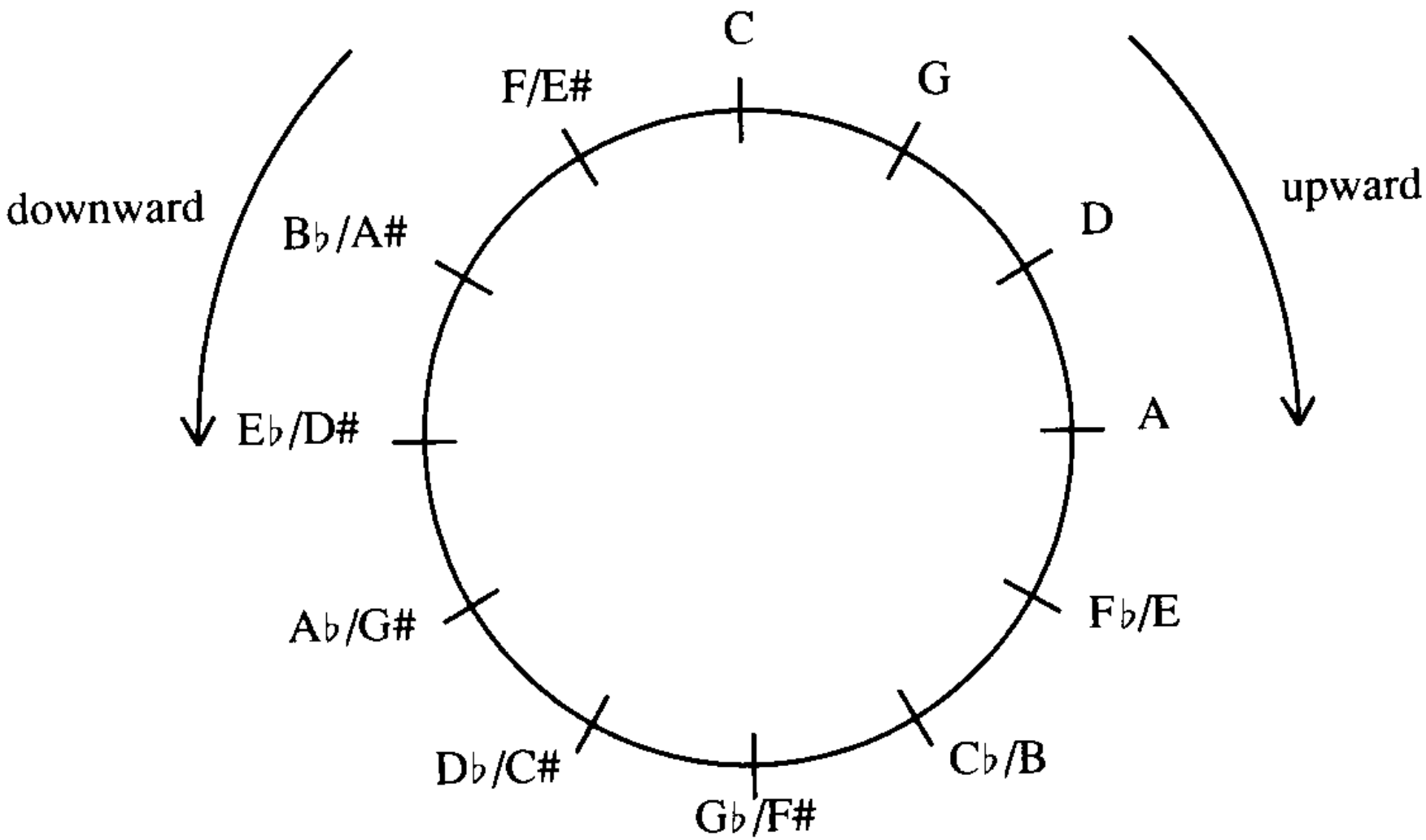


Pythagorean Tuning

- 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

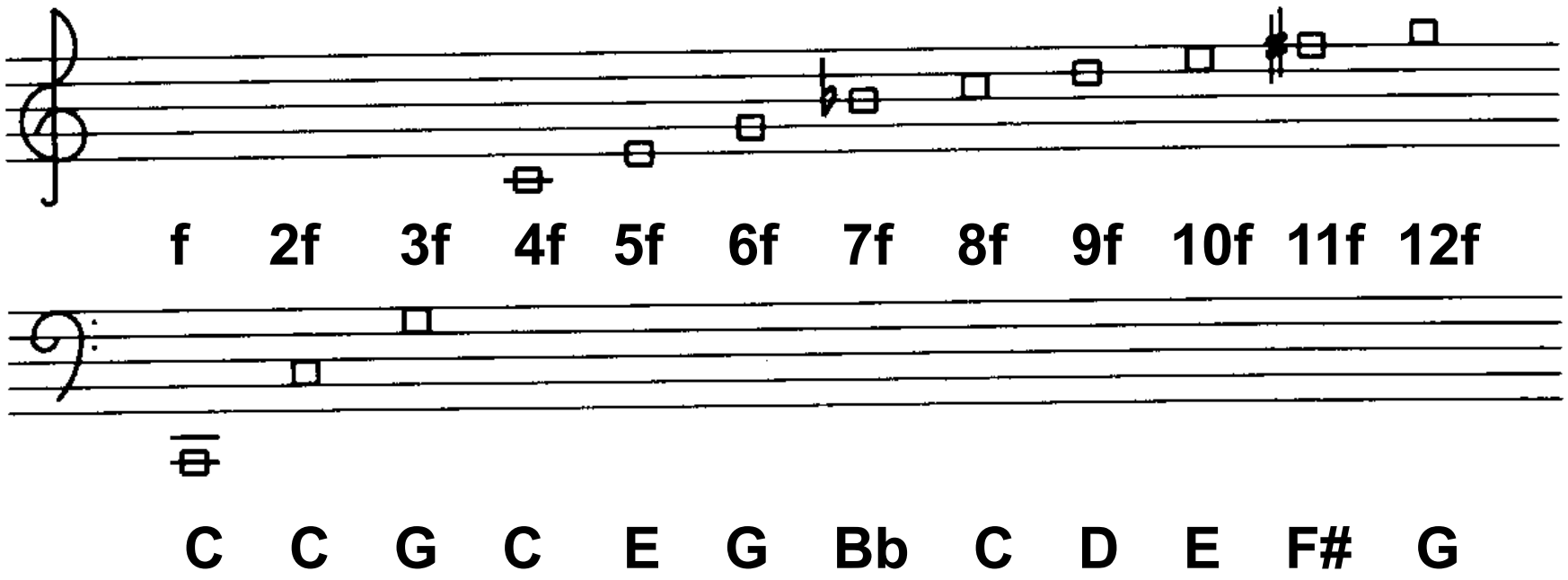
- What happens if you go around the circle of 5ths to get back to your original pitch class?

- $(3/2)^{12} = 129.75$
- 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



What is pitch?

- A perceptual attribute, 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, 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?



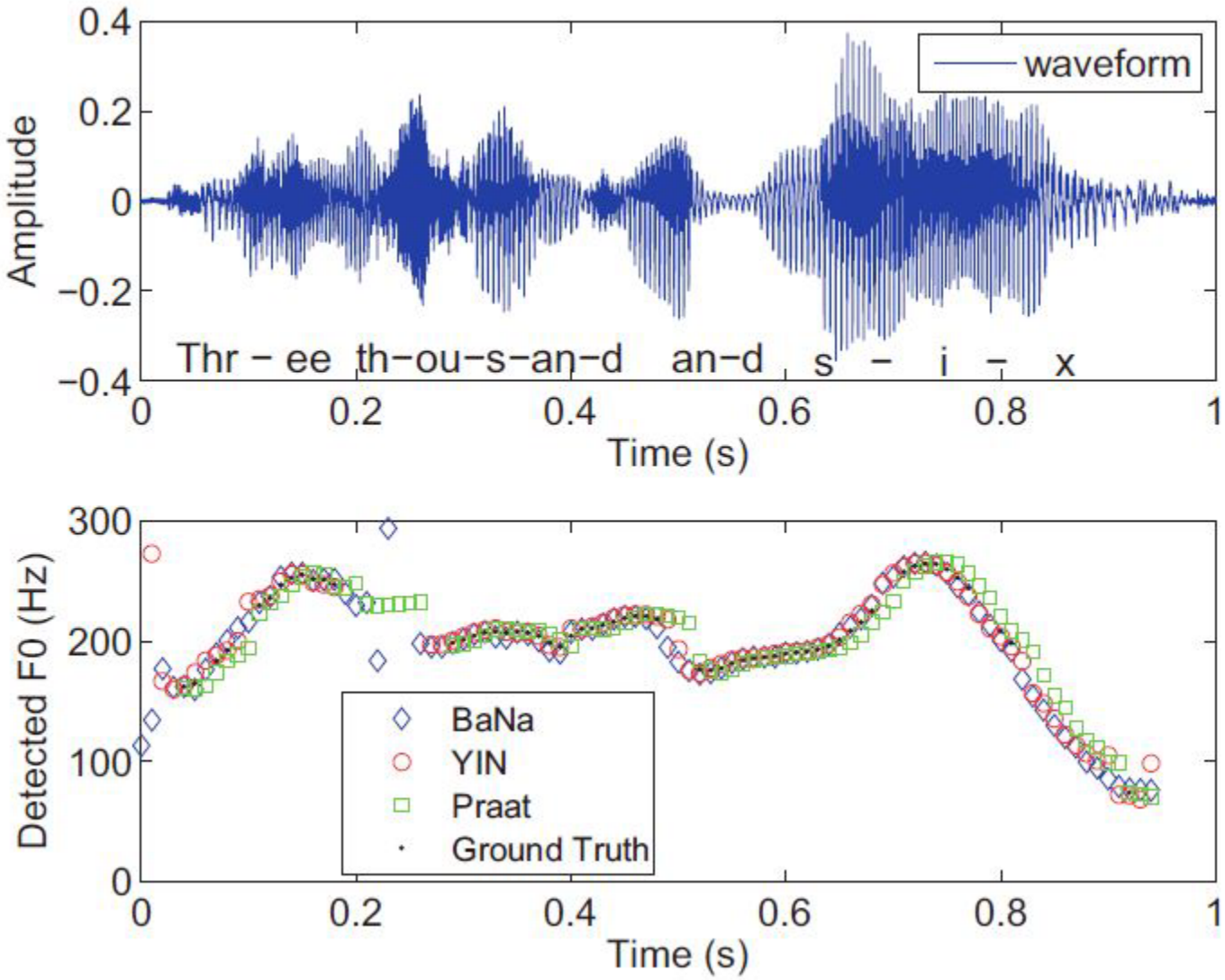


What emotion?

General Process of Pitch Detection

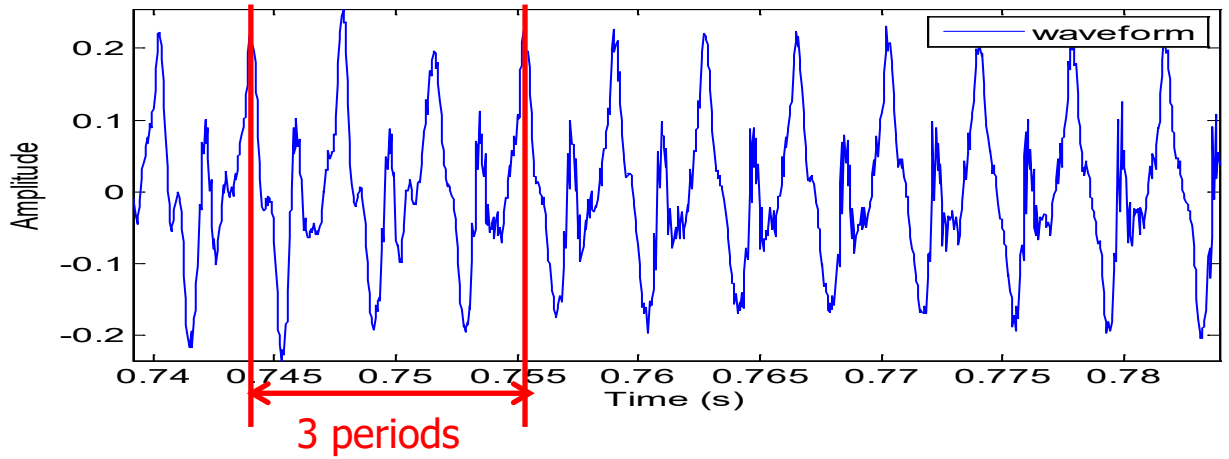
- Segment audio into time frames
  - Pitch changes over time
- Detect pitch (if any) in each frame
  - Need to detect if the frame contains pitch or not
- Post-processing to consider contextual info
  - Pitch contours are often continuous

An Example


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

How long should the frame be?

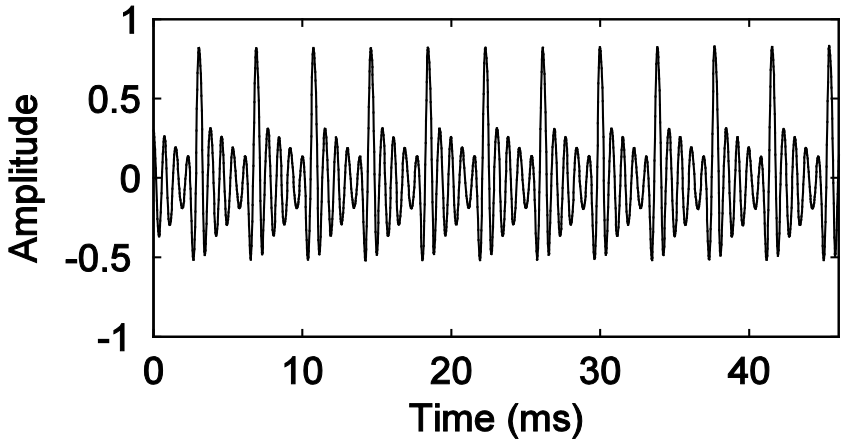
- Too long:
  - Contains multiple pitches (low time resolution)
- Too short
  - Can't obtain reliable detection (low freq resolution)
  - Should be at least about 2-3 periods of the signal

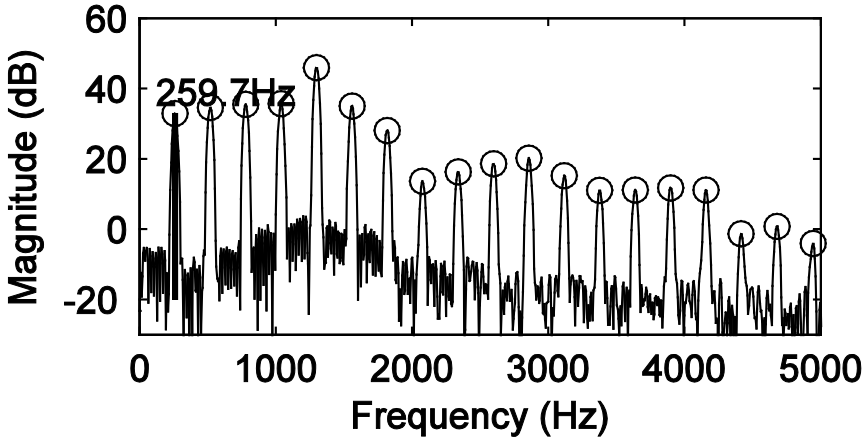


  - For speech or music, how long should the frame be?

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
- Spectral peaks have harmonic relations
  - F0 is the greatest common divisor
- Spectral peaks are equally spaced
  - F0 is the frequency gap

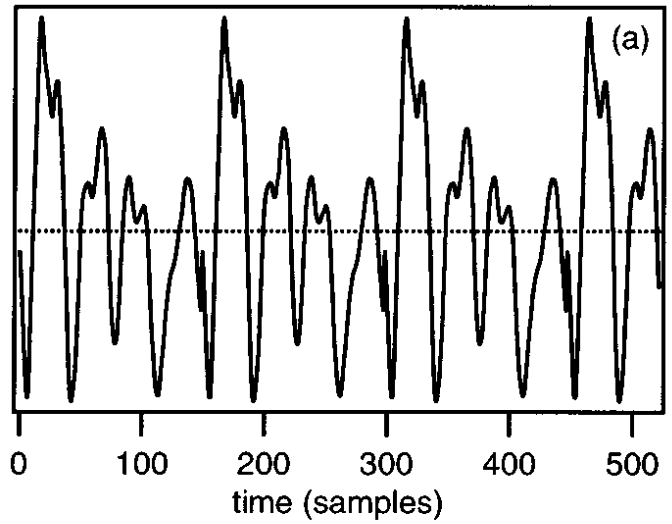
- Time domain
  - Detect period
- Frequency domain
  - Detect the divisor
- Cepstrum domain
  - Detect the gap

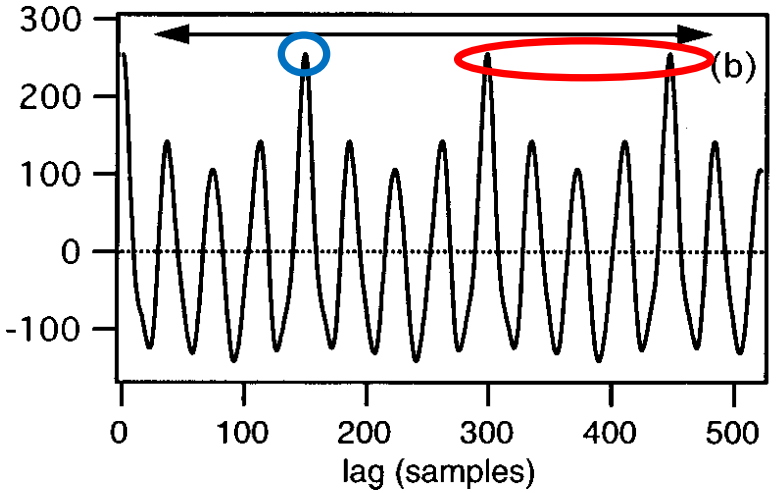
Time Domain: Autocorrelation

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

- 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)





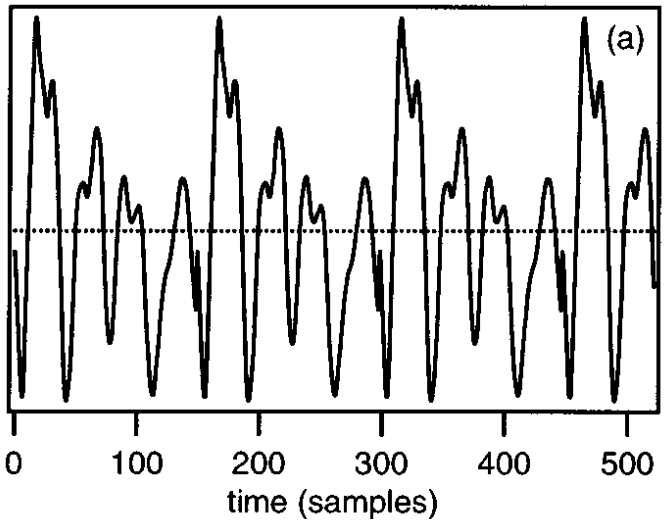
YIN – Step 2

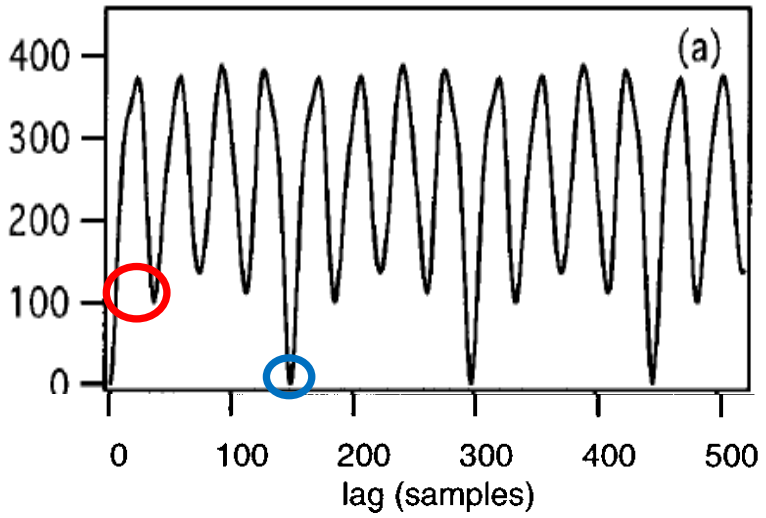
- Replace ACF with difference function

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

- Look for dips instead of peaks, which is why it's called YIN opposed to YANG.
- More immune to amplitude changes
- Problem
  - Some dips close to 0 lag might be deeper due to imperfect periodicity
  - Value scale is arbitrary, hard to set threshold

[de Cheveigne, 2002]





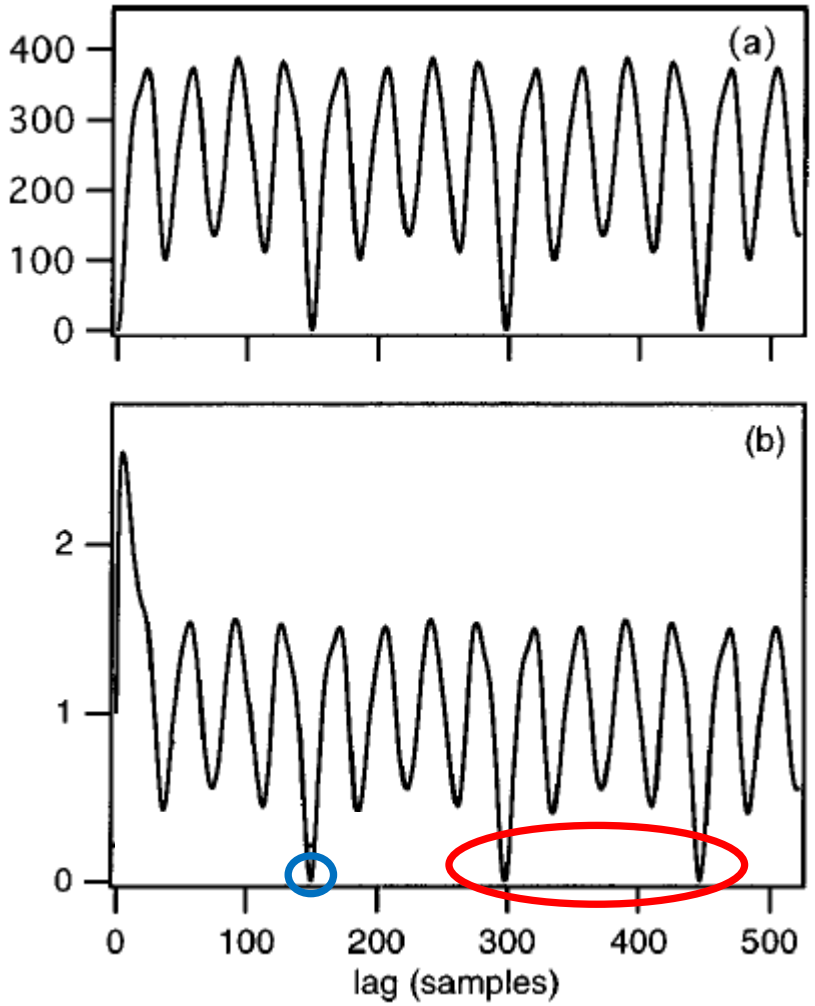
YIN – Step 3

- Cumulative mean normalized difference function

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

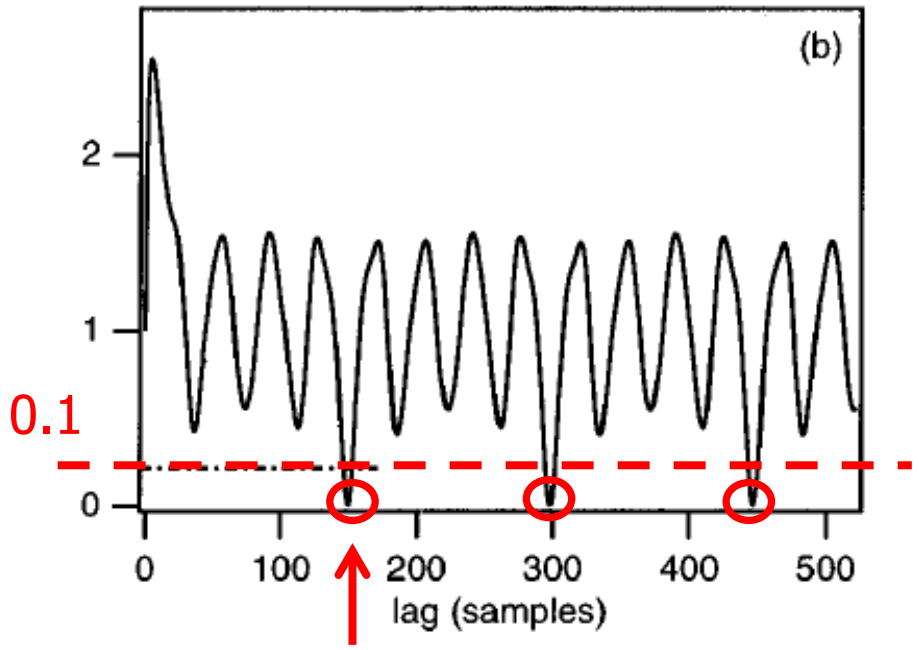
- Then take the deepest dip?

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



YIN – Step 4

- Absolute Threshold
  - Set threshold, e.g., 0.1
  - Pick the first dip below the threshold



YIN - Step 5

- Step 5: parabolic interpolation to find the exact dip location
  - The dip location in the discrete world may deviate from the exact dip location
    - For example, suppose sampling rate is 44,100 and the F0 is 990 Hz, then the period is 44100/990=44.55 samples. Our closest integer period estimate is 45 samples, corresponding to an F0 estimate of 44,100/45 = 980 Hz!
  - Use the three neighboring samples around the dip to perform a parabolic interpolation, and then find the dip of the parabola as the estimate of the period

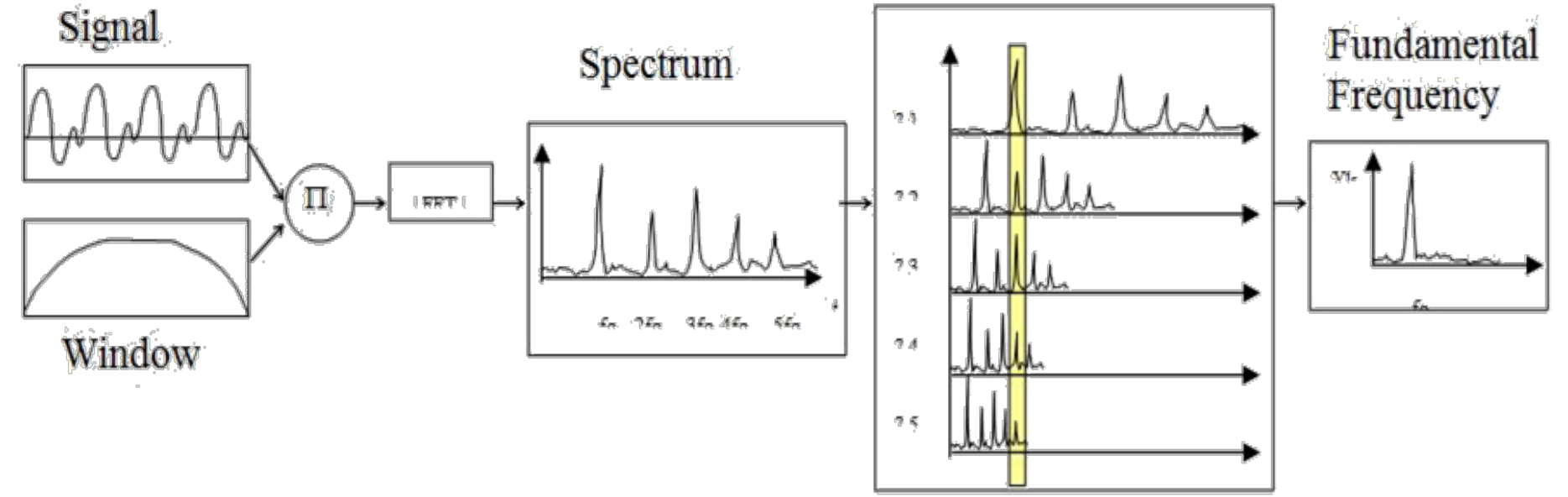
YIN – Step 6

- Step 6: use the best local estimate
  - The calculation of d and d’ requires a shorter window size than the frame size
  - This makes it possible to estimate d and d’ at different positions within the same frame
  - Some positions may be better than others (with smaller d’ values)
  - Use the pitch estimate from the best analysis position within the frame

Frequency Domain Approach

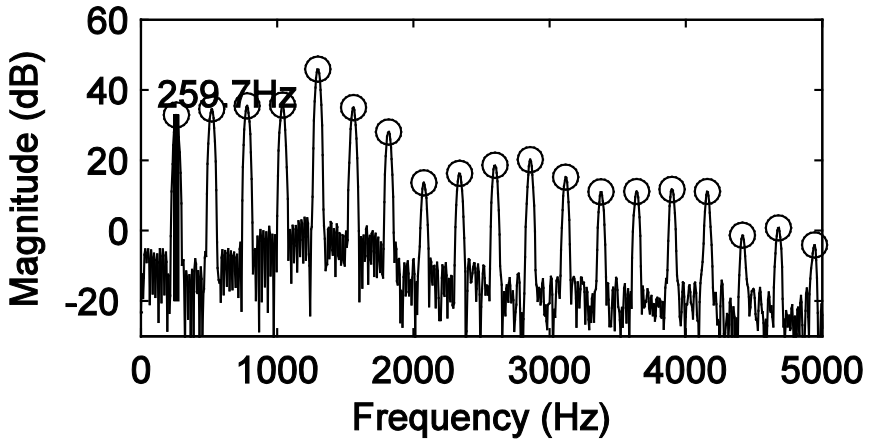
- Idea: for each F0 candidate, calculate the support (e.g., spectral energy) it receives from its harmonic positions.
- Harmonic Product Spectrum (HPS)

[Schroeder, 1968; Noll, 1970]



Cepstral Domain Approach

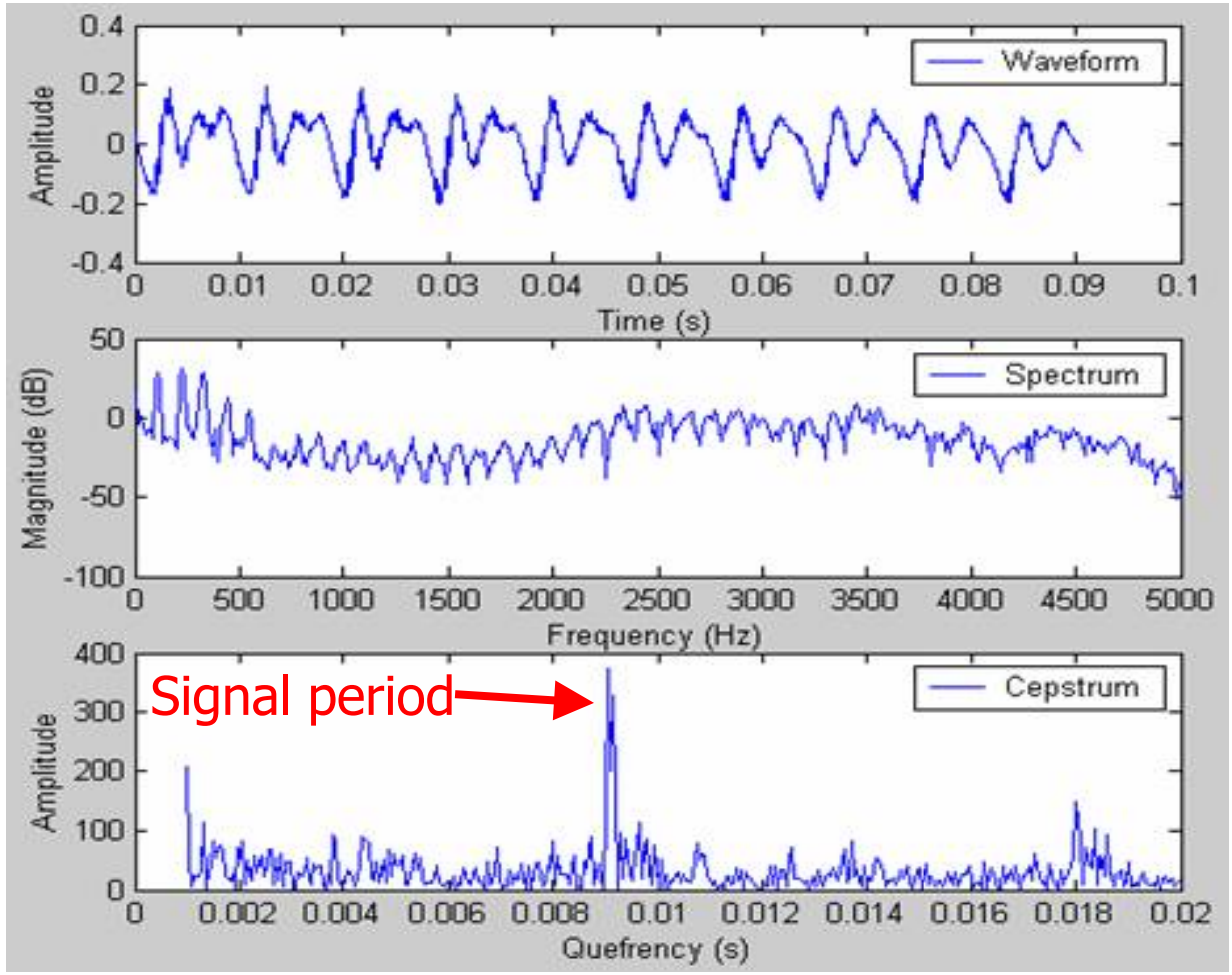
- Idea: find the frequency gap between adjacent spectral peaks
  - The log-amplitude spectrum looks kind of periodic
  - The gap can be viewed as the period of the spectrum
  - How to find the period then?
  - Cepstrum's idea: Fourier transform!



Cepstrum

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

Spectrum - Cepstrum
Frequency - Quefrency
Filtering - Liftering



Pitched or Non-pitched?

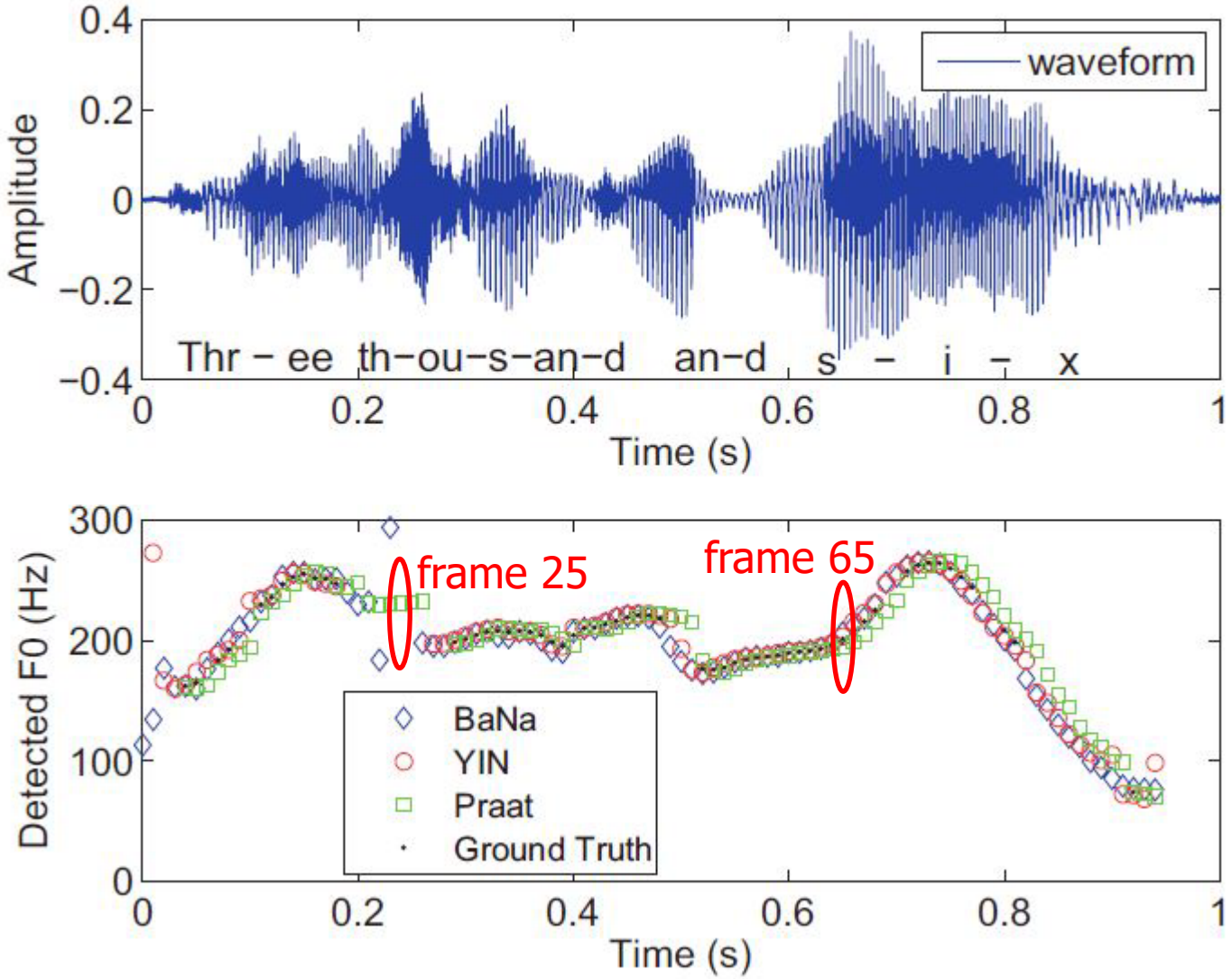
- Some frames may be silent or inharmonic, so they may not contain a pitch at all
  - Silence can be detected by RMS value
  - How about inharmonic frames?

- 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?

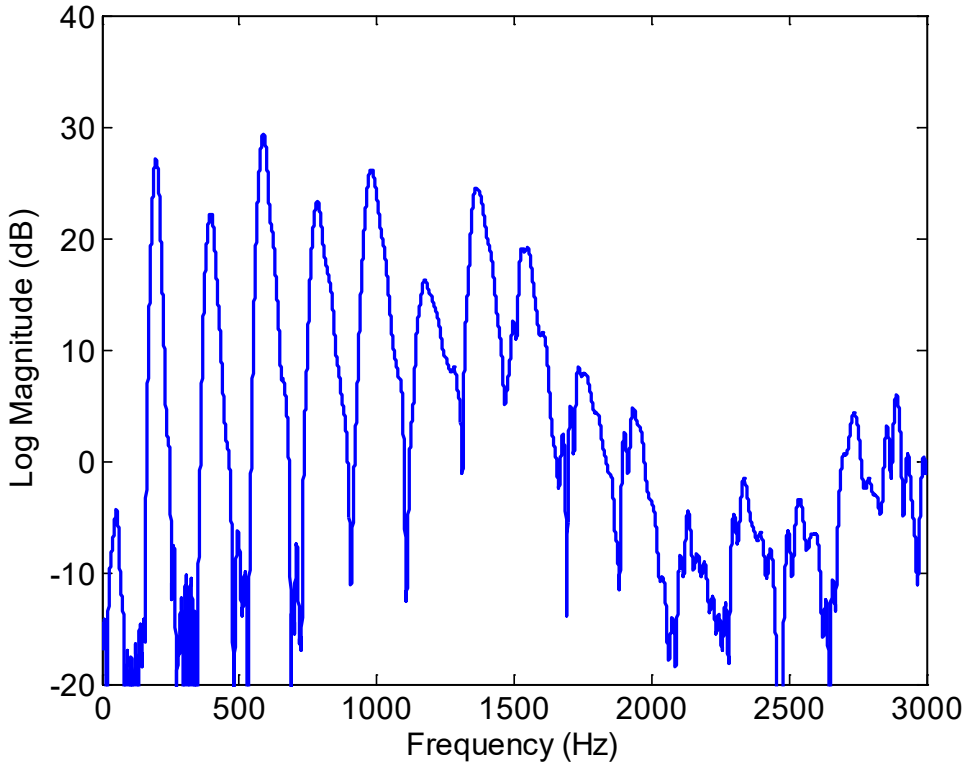
- Choose some recordings (speech, music)
- Get ground-truth
  - Listen to the signal and inspect the spectrum to manually annotate (time consuming!)
  - Automatic annotation using simultaneously recorded electrolaryngograph signals for speech (not quite reliable!)
- Pitched/non-pitched classification error
- Calculate the difference between estimated pitch with ground-truth
  - Threshold for speech pitch analysis: 10% or 20% in Hz
  - Threshold for music transcription: 1 quarter-tone (about 3% in Hz)

Different Methods vs. Ground-truth



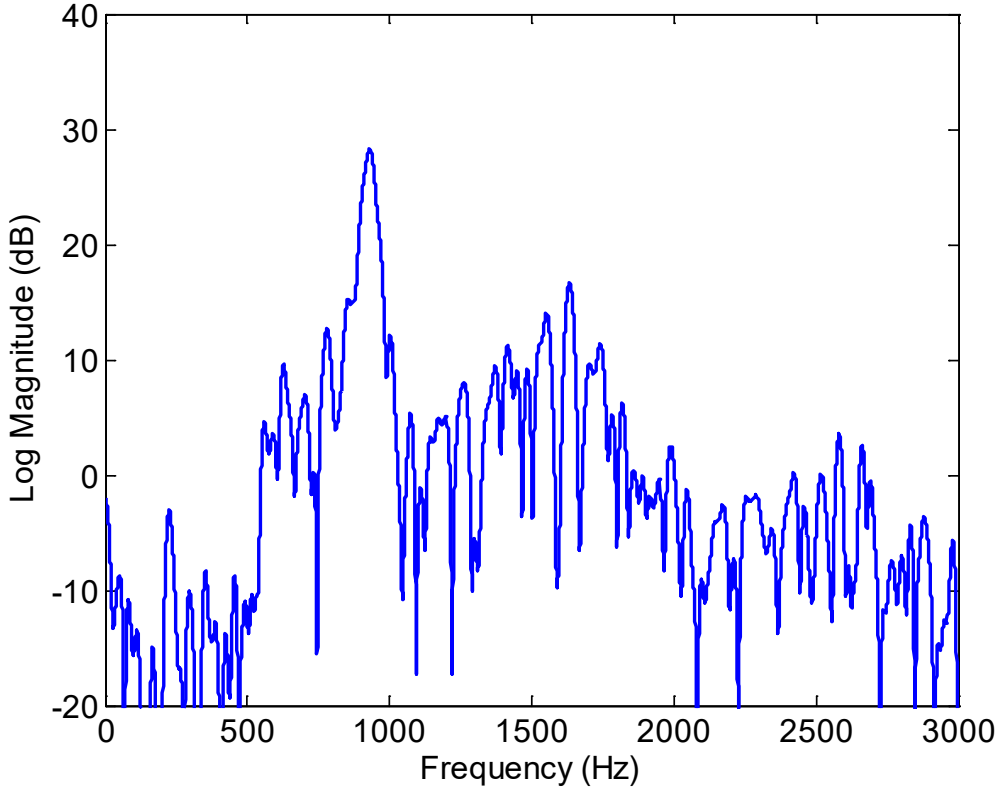
Frame 65 – Pitched (Voiced)

- 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?



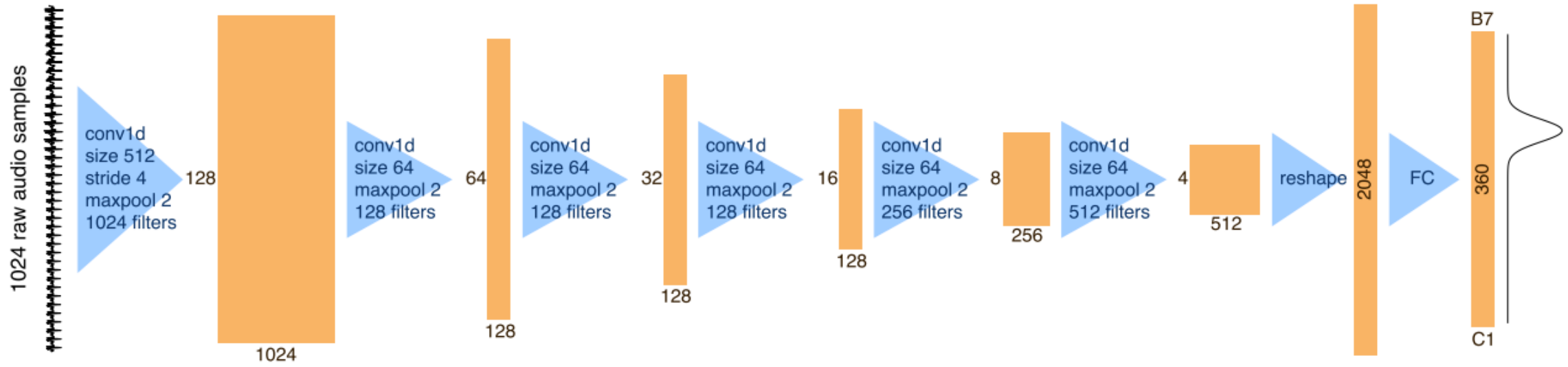
Violin + babble noise

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

Supervised Learning Method

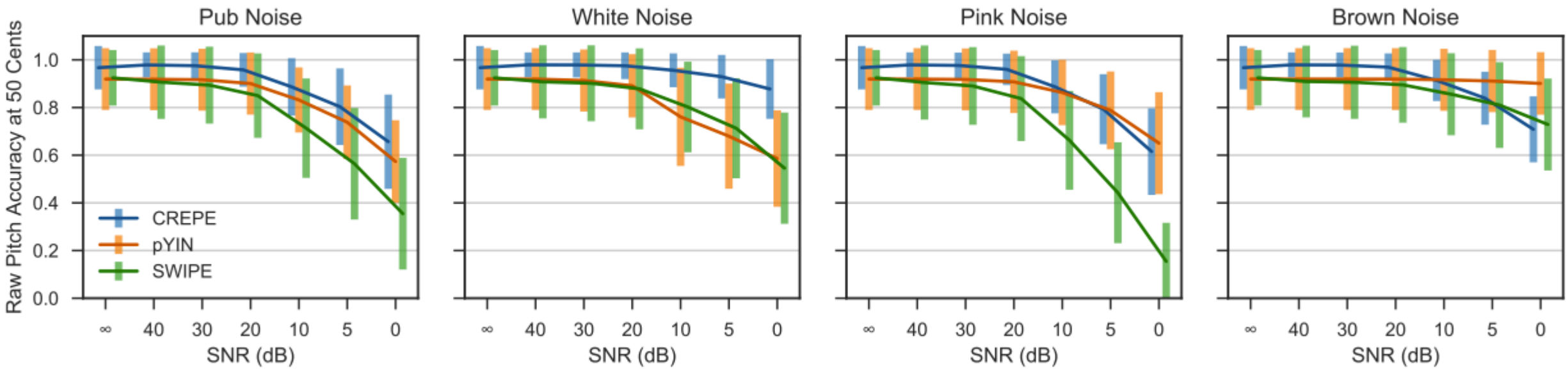
- Model input: audio frame; model output: target pitch
- Data driven; can be trained on specific type of data or diverse data

- 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

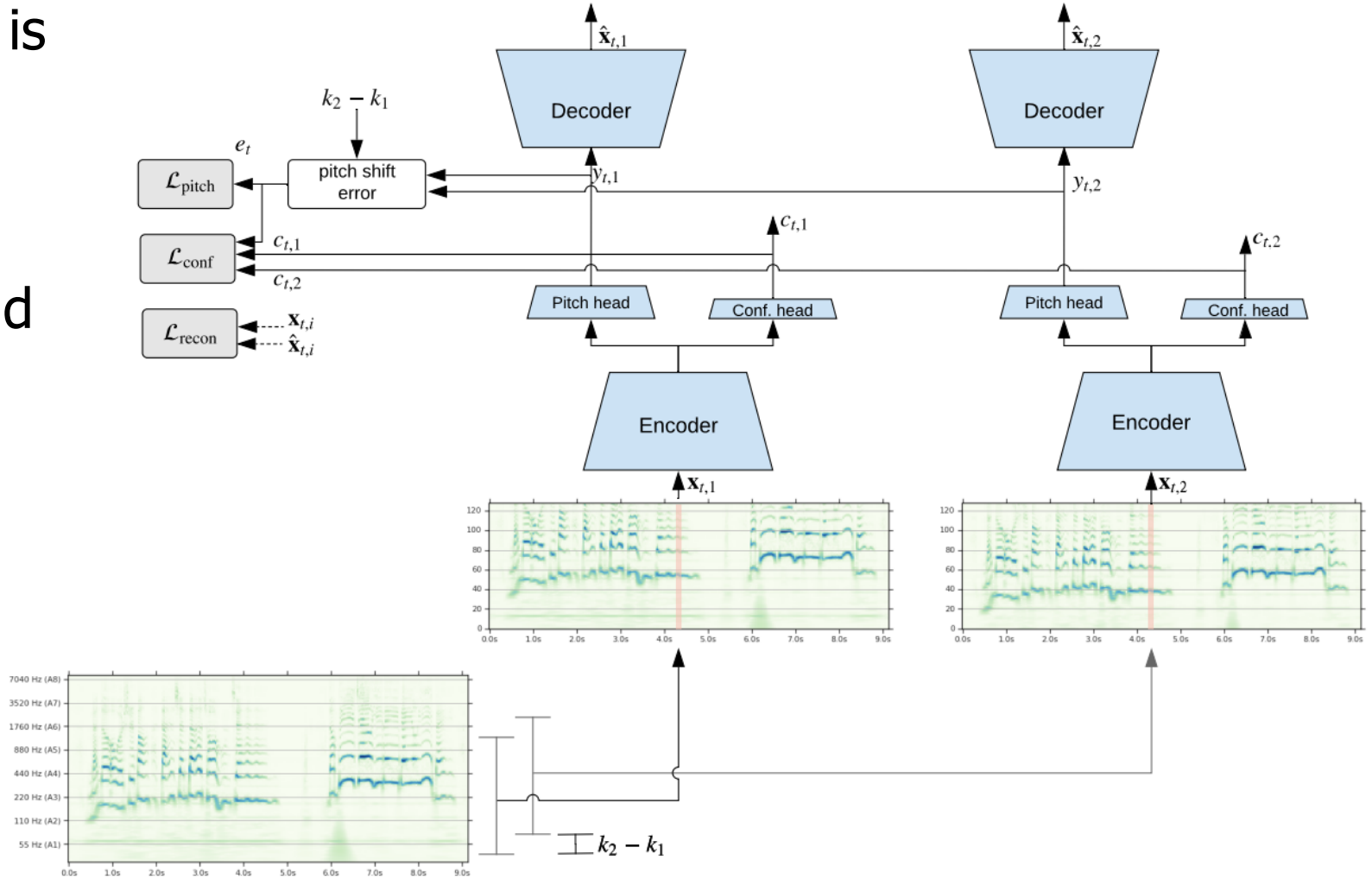


- Robustness can be further improved with data augmentation
- Online repo and model: https://github.com/marl/crepe

- 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

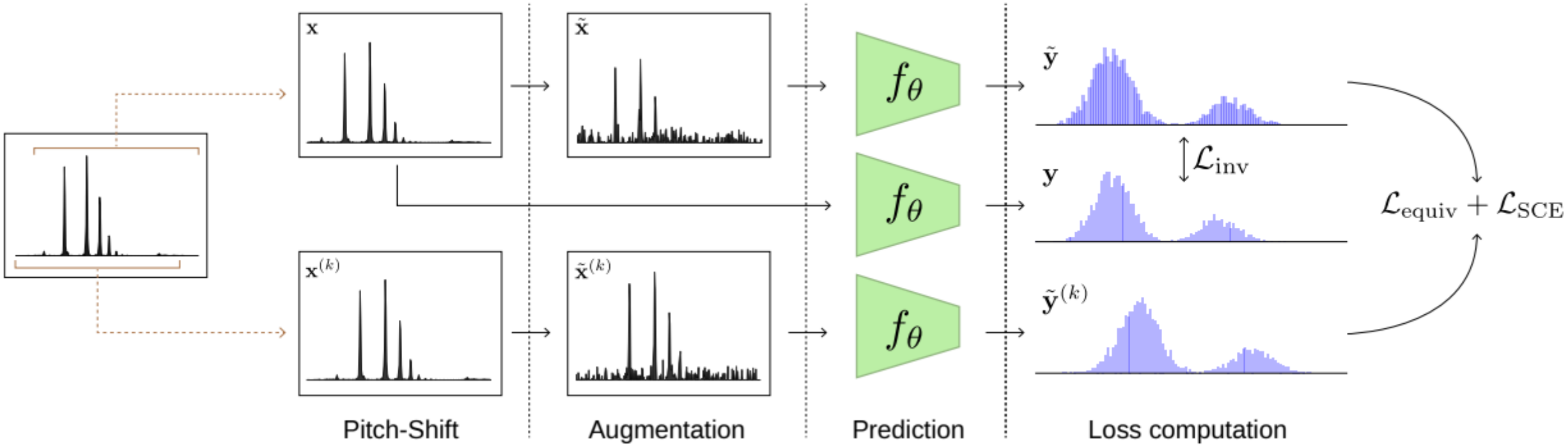
Model	# params	Trained on	*MIR-1k* RPA (CI 95%)	*MIR-1k* VRR	*MDB-stem-synth* RPA (CI 95%)
SWIPE	-	-	86.6%	-	90.7%
CREPE tiny	487k	many	90.7%	88.9%	93.1%
CREPE full	22.2M	many	90.1%	84.6%	92.7%
SPICE	2.38M	*SingingVoices*	$90.6\% \pm 0.1\%$	86.8%	$89.1\% \pm 0.4\%$
SPICE	180k	*SingingVoices*	$90.4\%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

Summary

- Pitch detection is important for many tasks
- 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

- Traditional signal processing methods
  - 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
- Moving from traditional signal processing to deep learning
  - HPS [Schroeder, 1968; Noll, 1970], Praat [Boersma, 1993], YIN [de Cheveigne, 2002], pYIN [Mauch & Dixon, 2014], BaNa [Yang et al., 2014]; PEFAC [Gonzales & Brookes, 2014]
  - Crepe [Kim et al., 2018]
- Self-Supervised Learning (SSL) shows promising results
  - SPICE [Gfeller et al., 2019], PESTO [Riou et al., 2023]

References

- Schroeder, M. R. (1968). Period histogram and product spectrum: New methods for fundamental frequency measurement. *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*.
- Childers, D. G., Skinner, D.P., and Kemerait, R.C. (1977). The cepstrum: A guide to processing. In *Proc. IEEE*.
- Paul Boersma, Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound, Institute of Phonetic Sciences, University of Amsterdam, Proceedings 17 (1993), 97-110.
- de Cheveigne, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. JASA.
- Mauch, Matthias, and Simon Dixon. "pYIN: A fundamental frequency estimator using probabilistic threshold distributions." in *Proc. IEEE International Conference on Acoustics, Speech And Signal Processing (ICASSP)*, pp. 659-663., 2014.
- 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.
- Alain Riou, Stefan Lattner, Gaëtan Hadjeres, and Geoffroy Peeters. "Pesto: Pitch estimation with self-supervised transposition-equivariant objective." In *Proc. International Society for Music Information Retrieval Conference (ISMIR),* 2023.