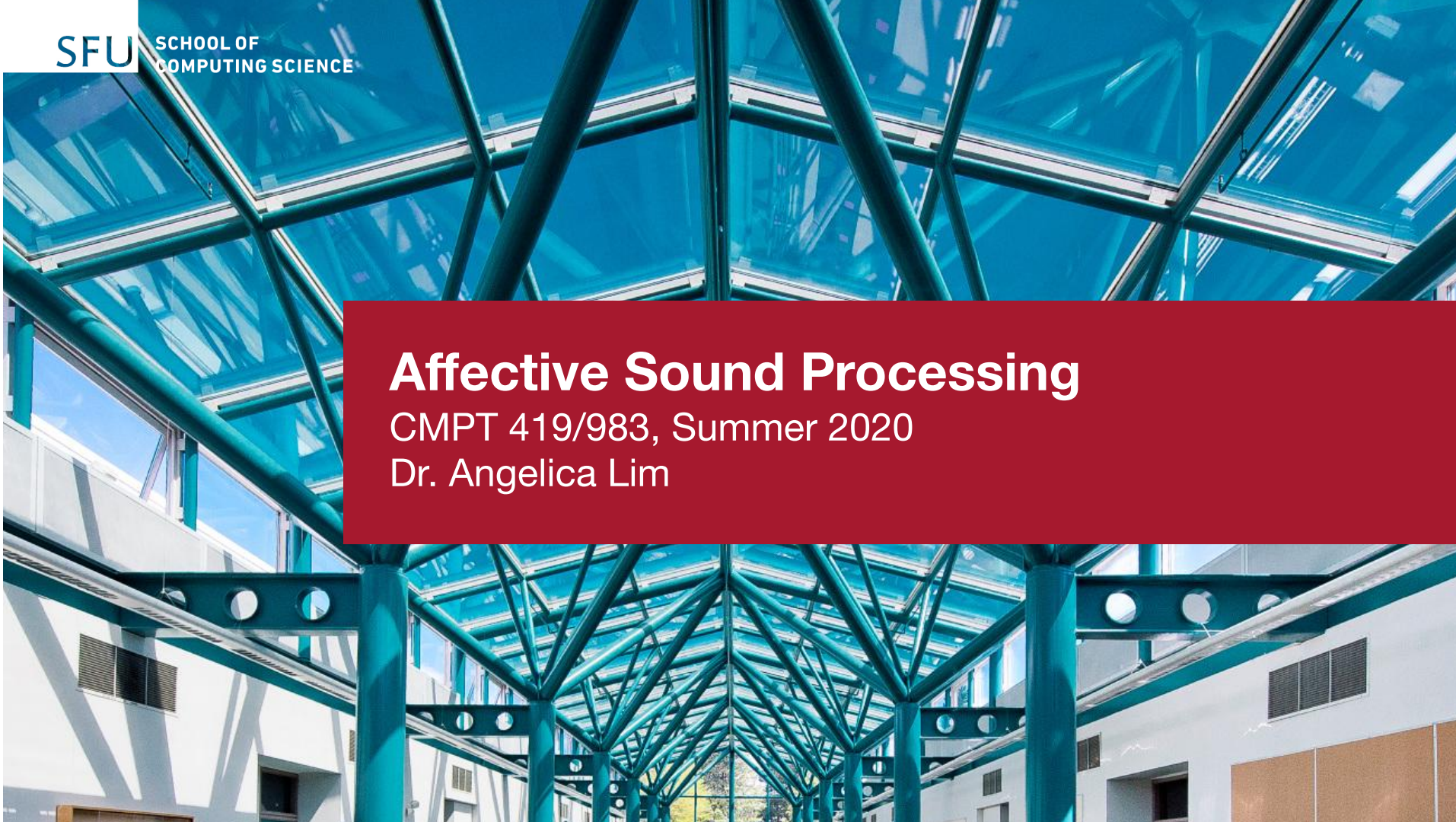


SFU SCHOOL OF COMPUTING SCIENCE

# Affective Sound Processing

CMPT 419/983, Summer 2020

Dr. Angelica Lim

# On Laughter

- More than jokes, happens in social interactions
- A story about a train...


Dr. Sophie Scott

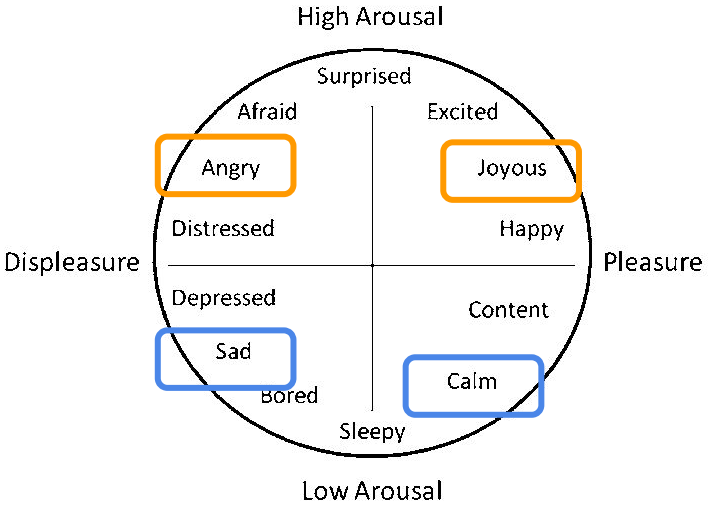


https://thepsychologist.bps.org.uk/volume-26/edition-4/laughter-ordinary-and-extraordinary

# Activity Debrief

Apply dynamic time warping to gesture data (A-Scary, B-Happy, C-Sad, D-Peaceful)

- Often people think of positive and negative emotions being opposite, but that's only in one dimension




Arousal is similar

E.g. "I did not expect happy and scary to be similar"

"Surprised that sad and peaceful were similar."

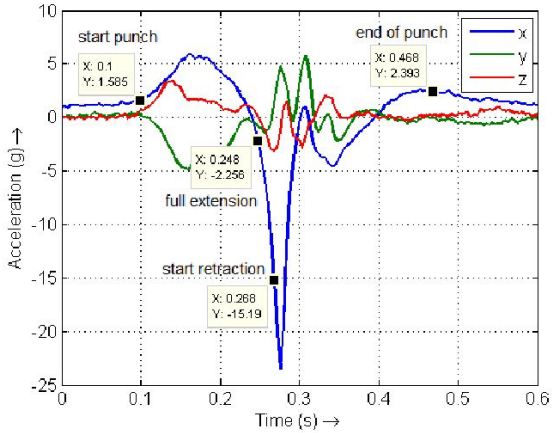
Arousal is well expressed in dynamics, whereas valence may be better expressed through face or other (e.g. major/minor scale in music)

# Activity Debrief

The difference between actions in multiple dimensions:

"I was expecting that DTW would present a small value between Knock and Wave for the same musical piece. For example, the distance between Knock_A and Wave_A would be closer than Knock_A and Wave_B. However, this did not happen consistently" - Nelson N.

Example of a jab



# Affect in Human Vocalizations

# Prosody




Vocalizations (Montreal affective voices):
Belin, P., Fillion-Bilodeau, S., & Gosselin, F. (2008). The Montreal affective voices: a validated set of nonverbal affect bursts for research on auditory affective processing. Behavior Research Methods, 40, 531-539.

Speech recording inventory:
Pell, M.D., Paulmann, S., Dara, C., Alasseri, A., & Kotz, S.A. (2009). Factors in the recognition of vocally expressed emotions: a comparison of four languages. Journal of Phonetics, 37, 417-435.

# Vocal features

**Table 6.1**

Summary of human vocal effects most commonly associated with the emotions indicated. Descriptions are given relative to neutral speech. (Adapted with permission from Murray and Arnott (1993), Table 1. Copyright 1993 Acoustical Society of America.)

| | Fear | Anger | Sadness | Happiness | Disgust |
|---|---|---|---|---|---|
| **Speech rate** | much faster | slightly faster | slightly slower | faster or slower | very much slower |
| **Pitch average** | very much higher | very much higher | slightly lower | much higher | very much lower |
| **Pitch range** | much wider | much wider | slightly narrower | much wider | slightly wider |
| **Intensity** | normal | higher | lower | higher | lower |
| **Voice quality** | irregular voicing | breathy chest tone | resonant | breathy blaring | grumbled chest tone |
| **Pitch changes** | normal | abrupt on stressed syllables | downward inflections | smooth upward inflections | wide downward terminal inflections |
| **Articulation** | precise | tense | slurring | normal | normal |

# Not *what* you say, but *how*

Applications of vocal affect recognition

- Irate people on the phone [annoyed/angry]
- Sexy voices [desire]
- Happy voices [amusement, contentment, joy]
- Stressed voice [stressed, worried, fearful]

Speech processing throws out all prosody:

Ok! :)  Ok.  Okaaay….?  Ok :(  Okaaayyyy :D

→ Off-the-shelf speech recognition systems cannot tell you if someone is asking a question (good project!)

# Affect Bursts

**Affect bursts**: short, emotional non-speech expressions




Experimental study of affect bursts (Shroder, 2003)

# This week

**Today:**

- Time Domain
- Energy
- Frequency Domain
- Spectrogram
- Pitch detection
- Fundamental frequency

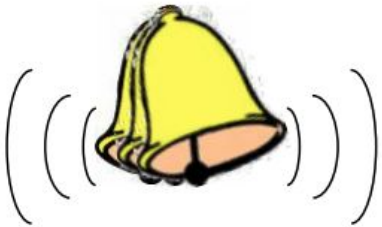
**Thursday**:

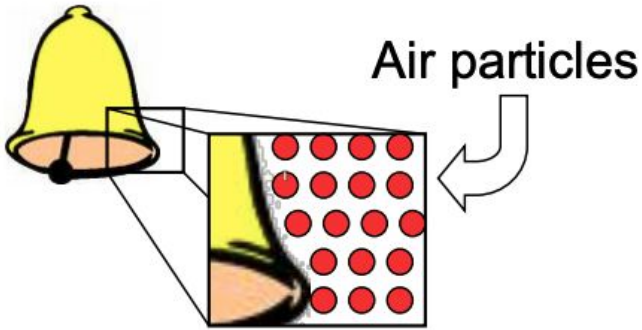
- Cepstrum
- MFCCs
- Formants

# What is sound?

Sound is produced by a vibrating source that causes the air around it to move.

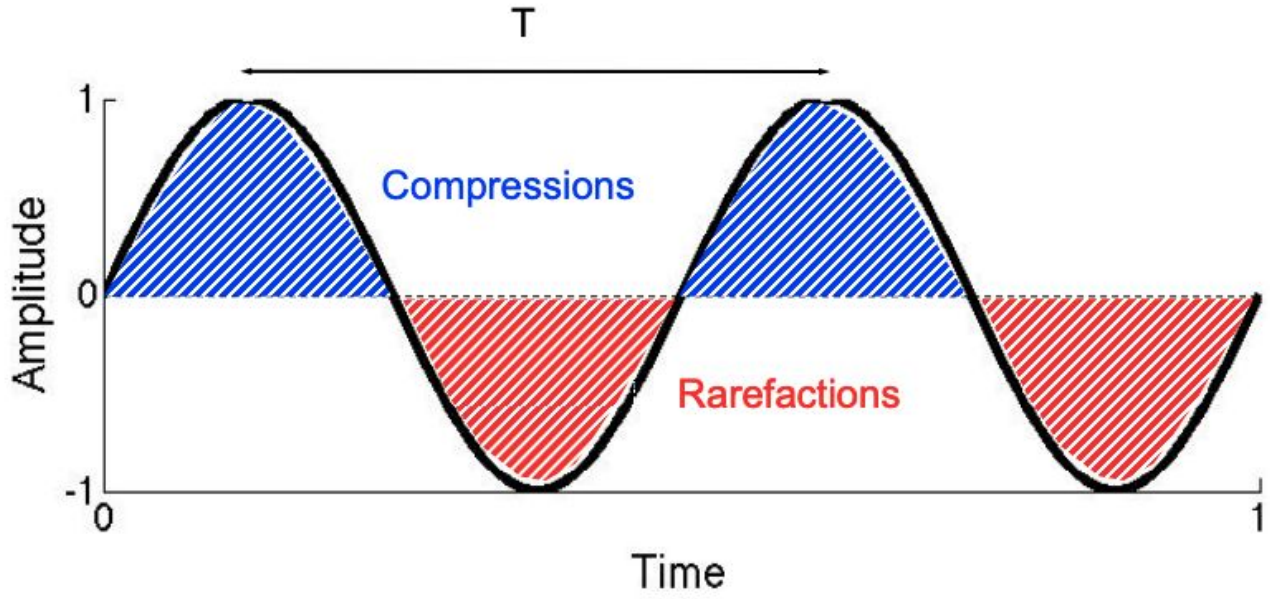
Pushes increase the air pressure, while pulls decrease the air pressure. The vibration sends a wave of pressure fluctuation through the air





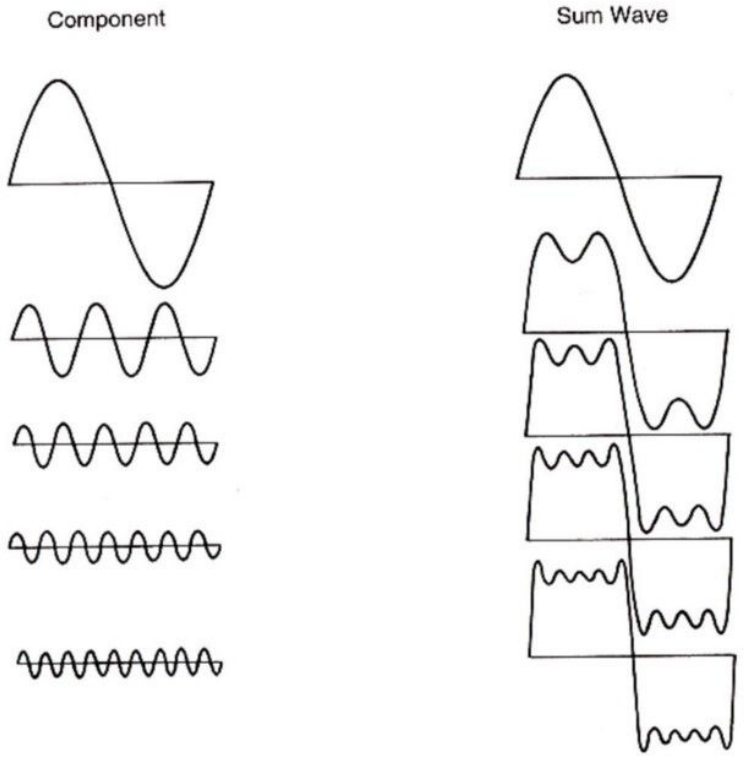
# Sound waves

$$x(t) = A \cdot sin(2\pi ft + \theta)$$

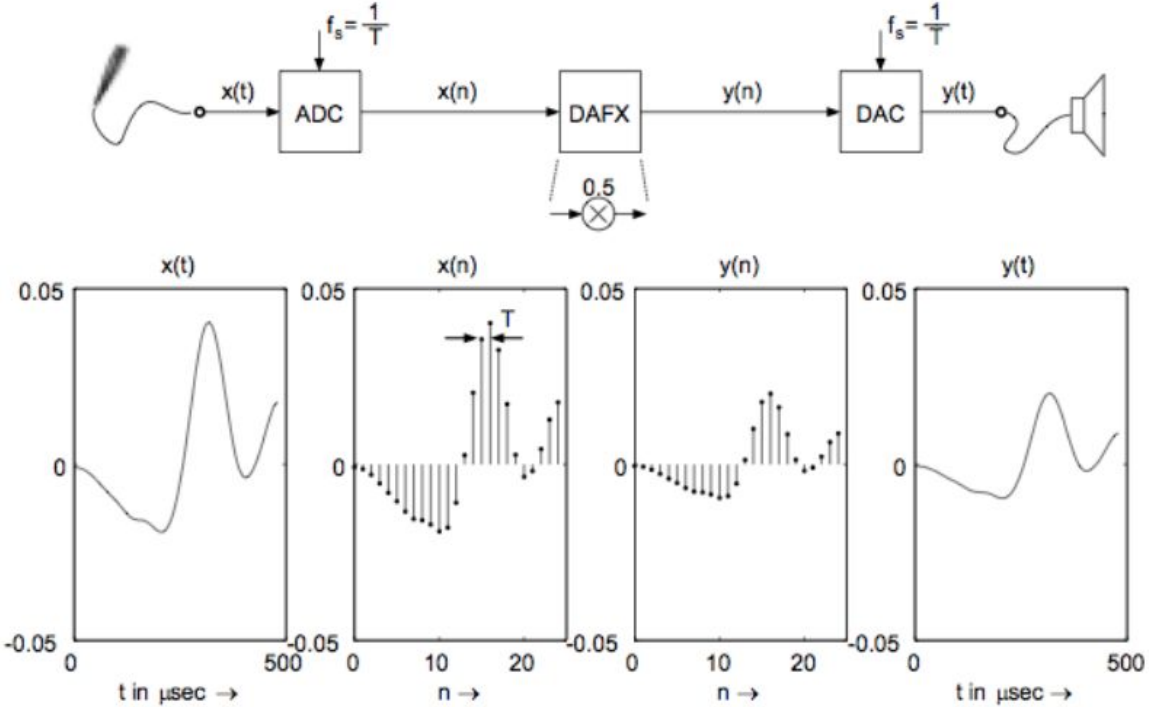


# Additive Waves

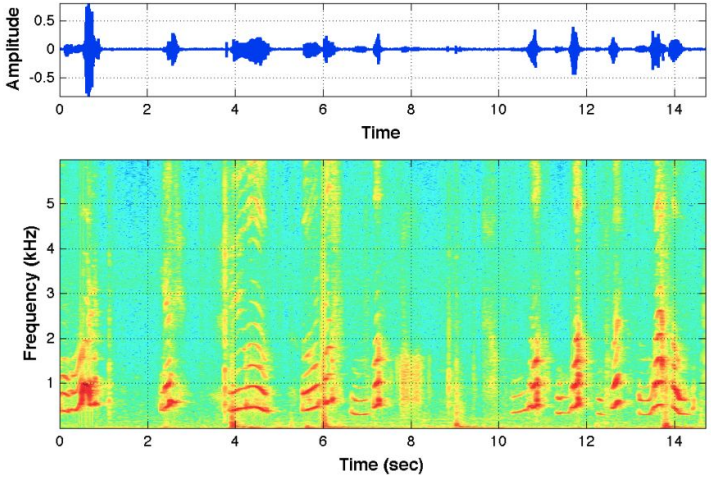
Simple sine waves can sum up to create more complex waves.

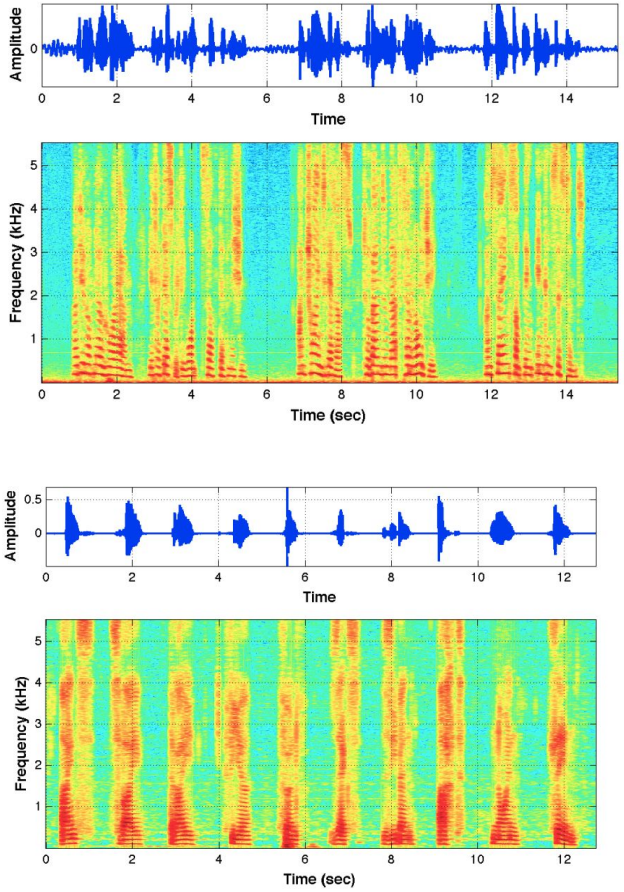


# Digital Audio



# Spectrogram



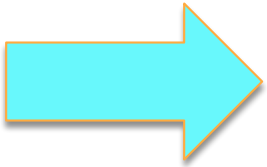



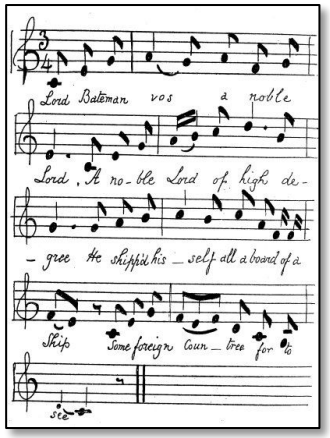
# Pitch Detection

*Based on Bello's Pitch Estimation Tutorial*

# Pitch Detection: The Problem

- Extract pitches from recordings or live music / vocal interactions

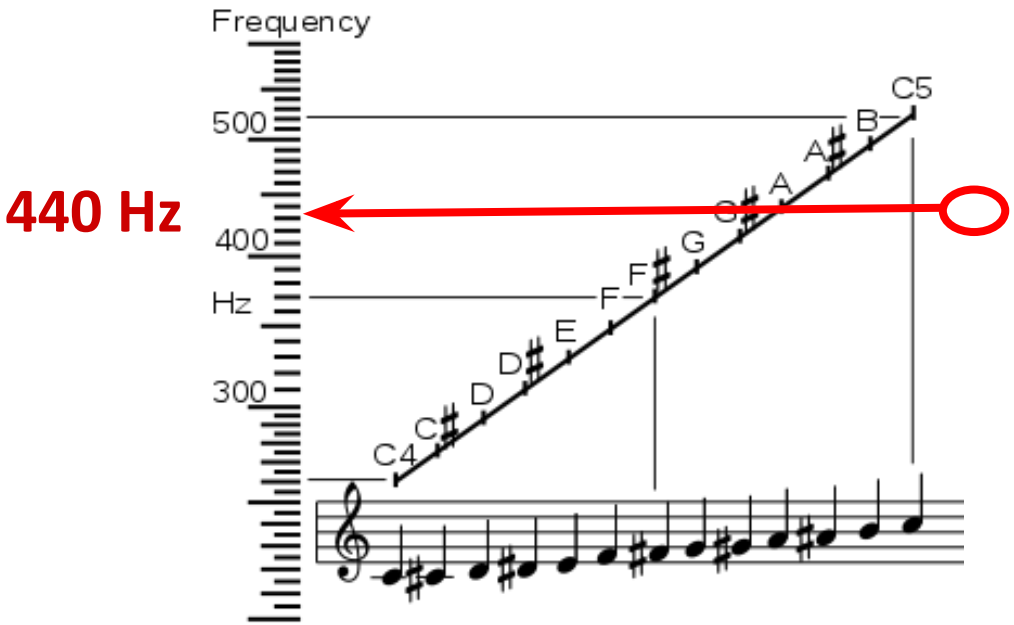


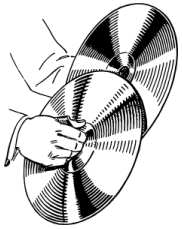



- Applications: music retrieval, speech analysis, automatic score transcription, score following (musical karaoke), and affective computing, of course!

# What is Pitch?

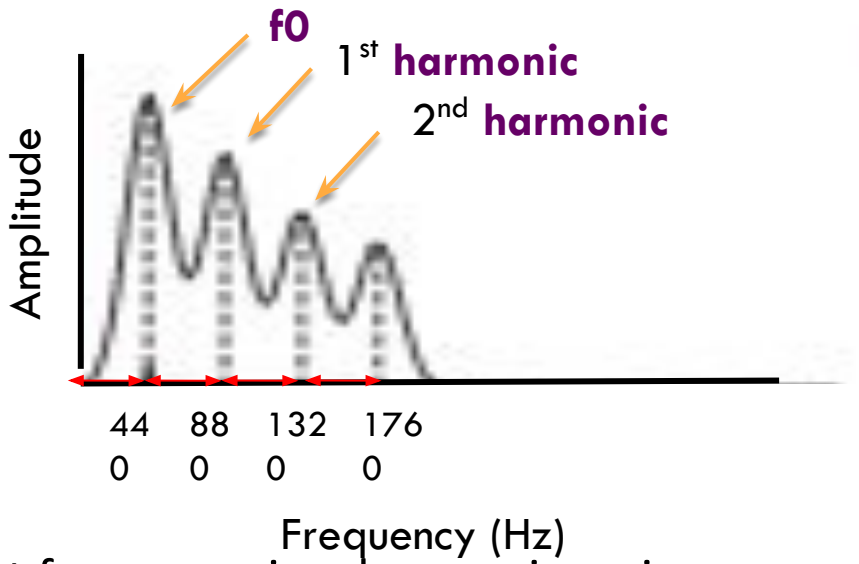
- Pitches correspond to frequencies



# Useful Definitions

Note: A (440 Hz) 🔊 → FFT →

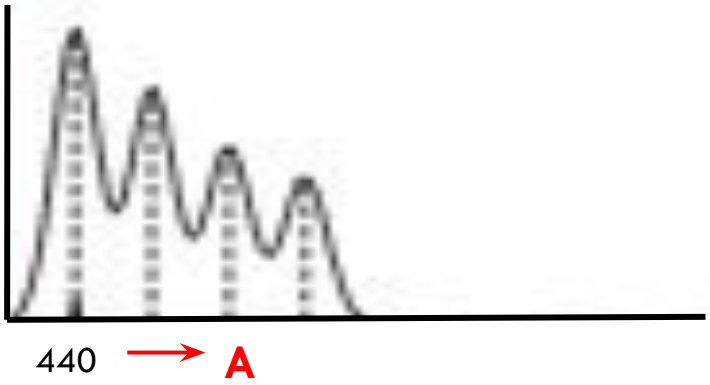



- **Fundamental Frequency (F0)**: Lowest frequency in a harmonic series
- **Harmonics**: Integer multiples of a frequency

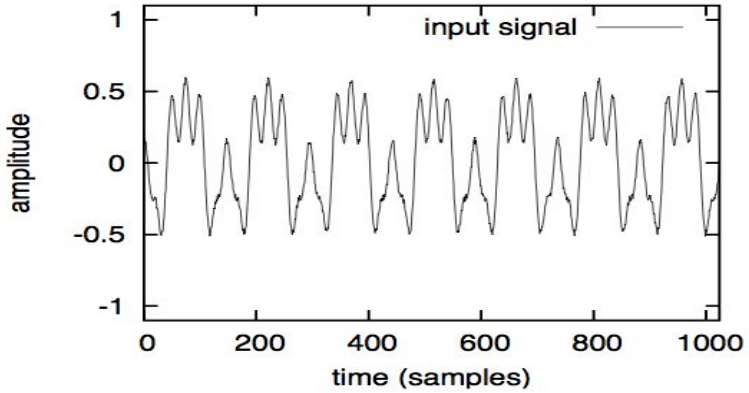
# What is Pitch Detection?

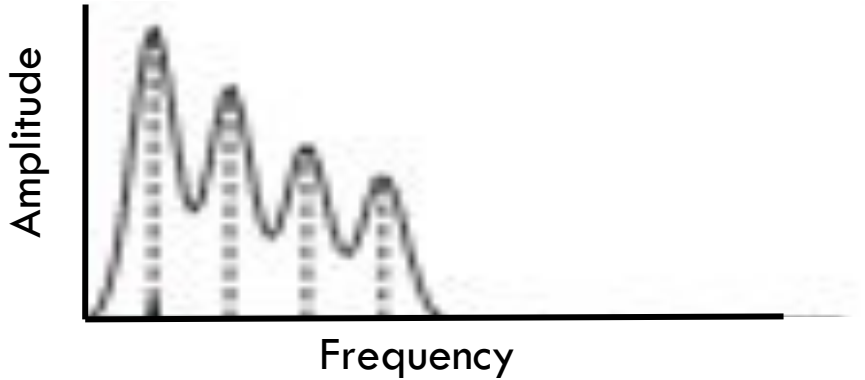
Typical Definition:

Pitch Detection = Fundamental Frequency (F0) Estimation



# Pitch Detection Techniques




- Time Domain
  - Zero crossing rate (ZCR)
  - Autocorrelation
    - Maximum product
    - YIN (Minimum difference)
- Frequency Domain
  - Autocorrelation
    - Spectral Comb
  - Harmonic product spectrum
  - Maximum Likelihood

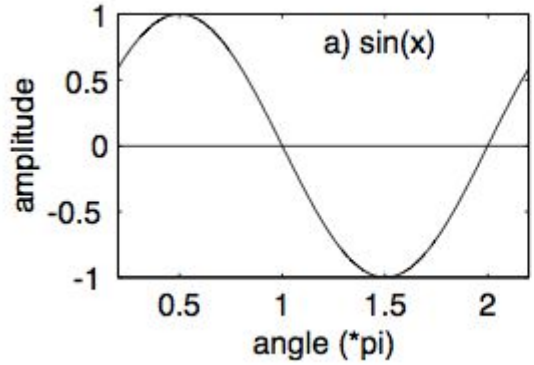
# Time Domain

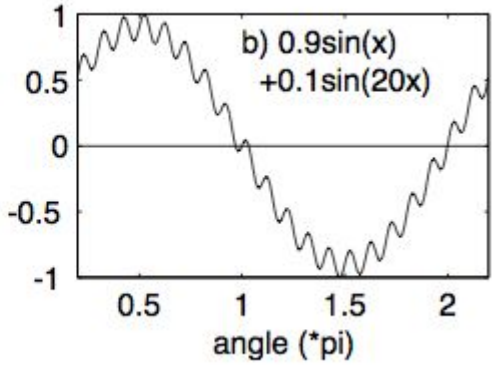
*Zero Crossings*

*Auto Correlation: Maximum Product*

*Auto Correlation: YIN (Minimum Difference)*

# Time Domain: Zero Crossing Rate



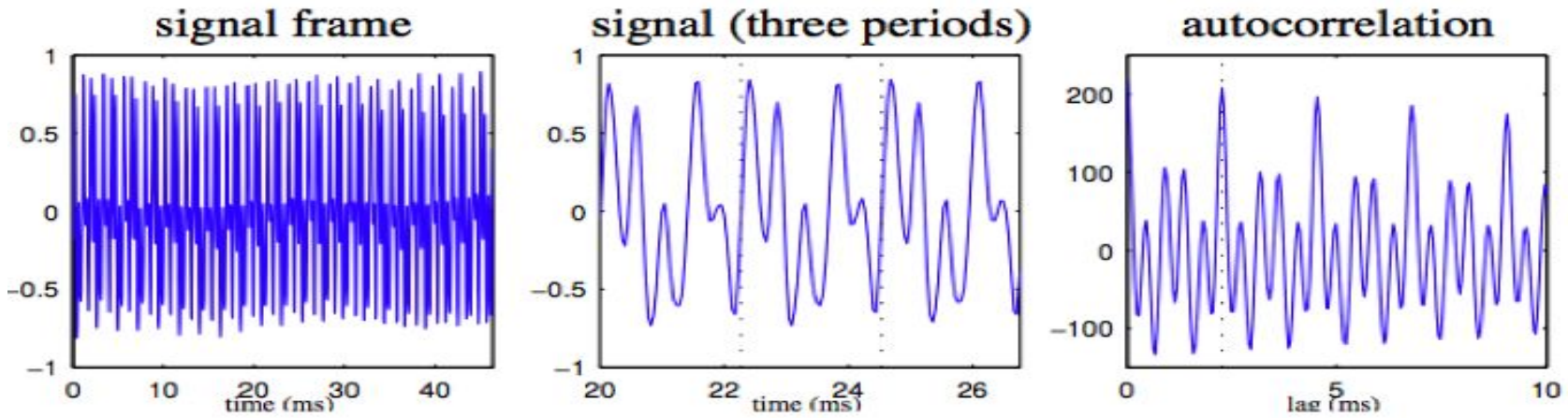




Bad

- Eg. 440 Hz = 440 cycles/sec

  = 880 zero crossings/sec

# Time Domain: Autocorrelation (AC)



# Time Domain: Autocorrelation (AC)

- Autocorrelation Function

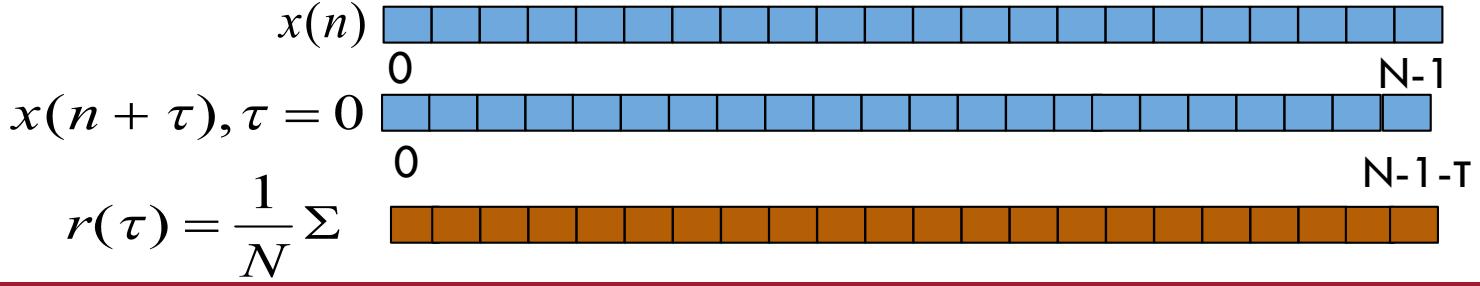
Find the maximum of this signal

lag

$$r(\tau) = \frac{1}{N} \sum_{n=0}^{N-\tau-1} x(n)x(n+\tau)$$

Normalize by # samples in x(n) (optional)

Input signal x(n)



# Time Domain: Autocorrelation (AC)

- Autocorrelation Function

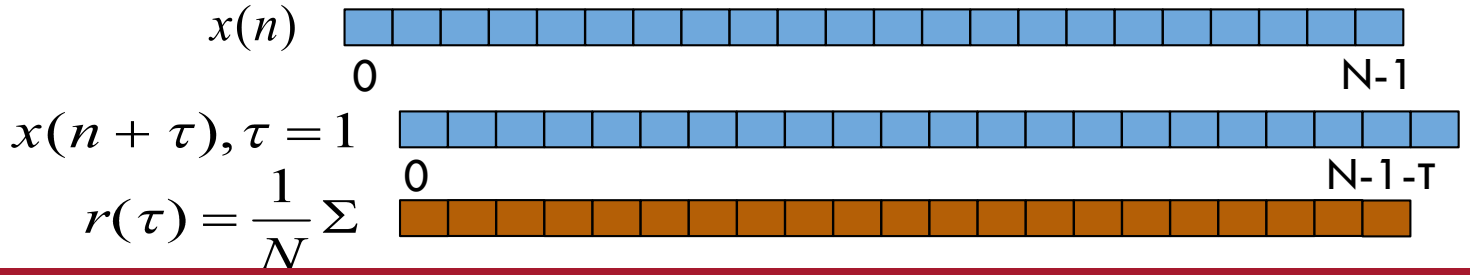
Find the maximum of this signal

lag

$$r(\tau) = \frac{1}{N} \sum_{n=0}^{N-\tau-1} x(n)x(n+\tau)$$

Normalize by # samples in x(n) (optional)

Input signal x(n)



# Time Domain: Autocorrelation (AC)

- Autocorrelation Function

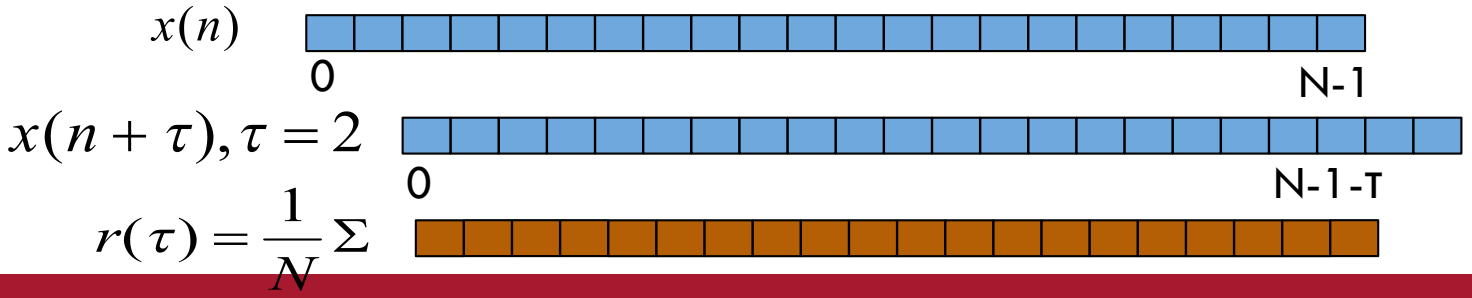
Find the maximum of this signal

lag

$$r(\tau) = \frac{1}{N} \sum_{n=0}^{N-\tau-1} x(n)x(n+\tau)$$

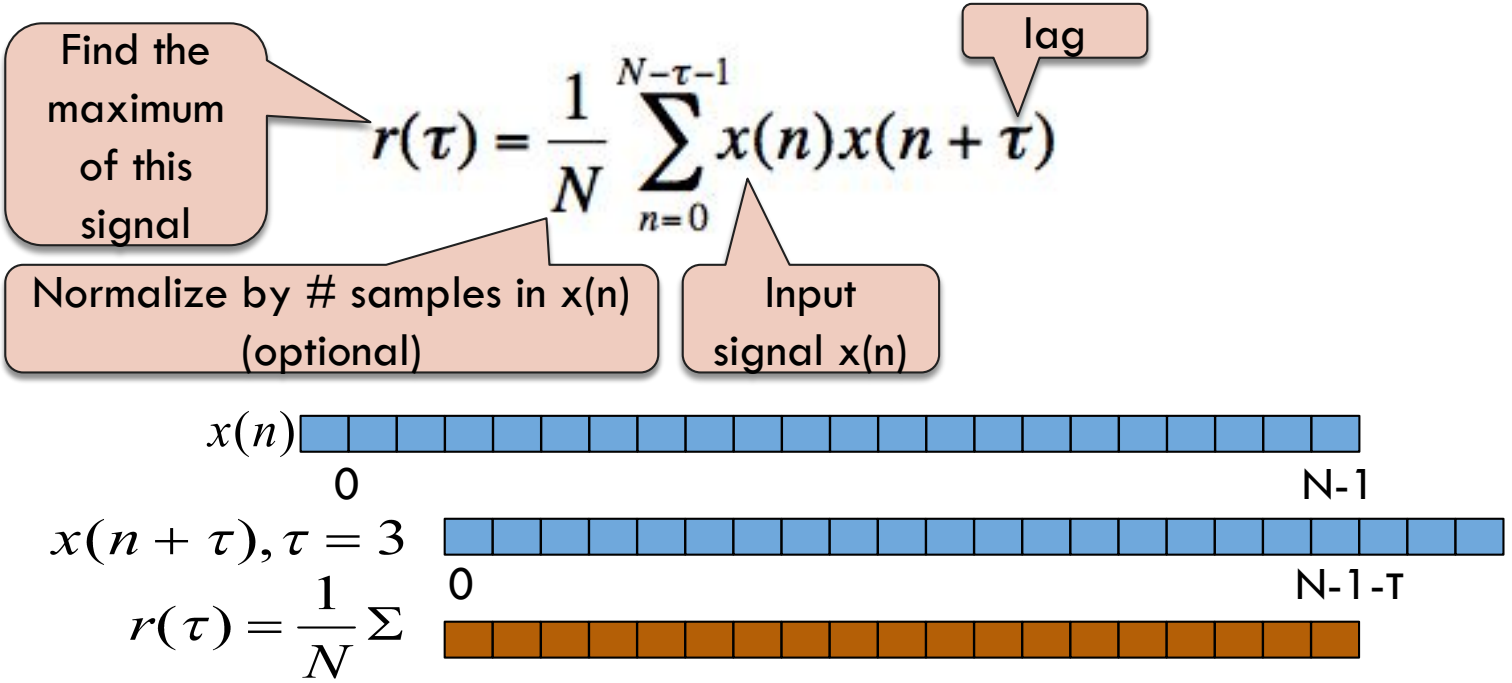
Normalize by # samples in x(n) (optional)

Input signal x(n)



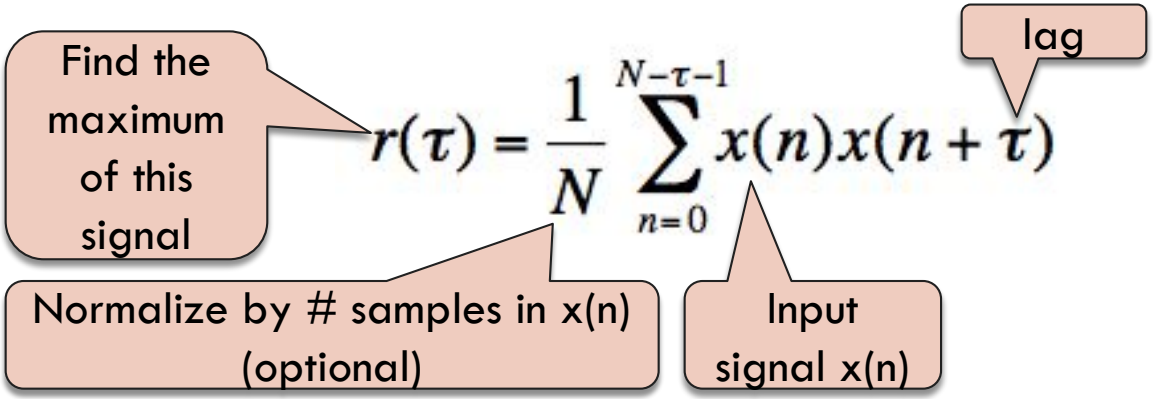
# Time Domain: Autocorrelation (AC)

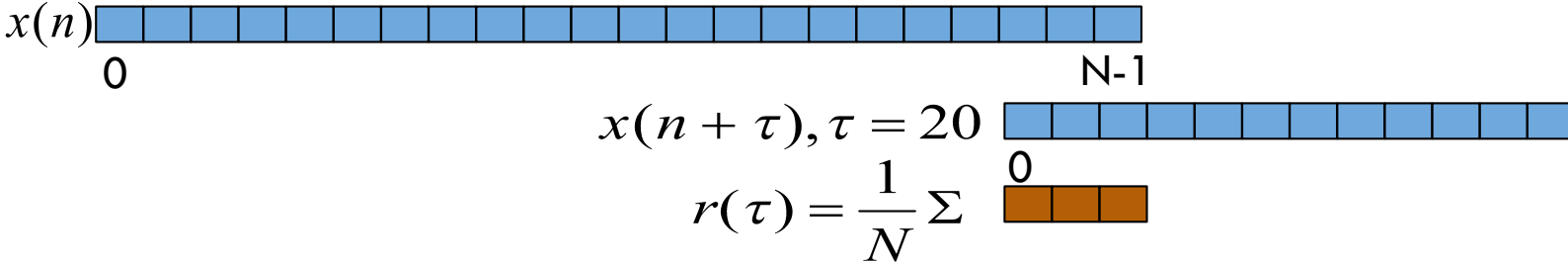
- Autocorrelation Function



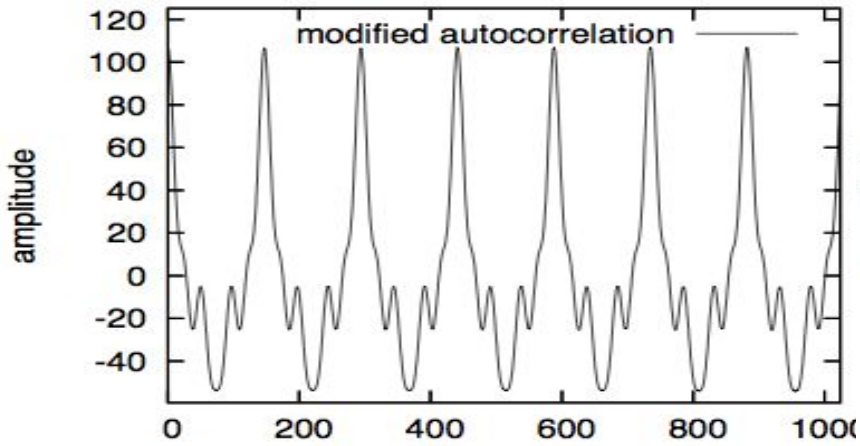
# Time Domain: Autocorrelation (AC)

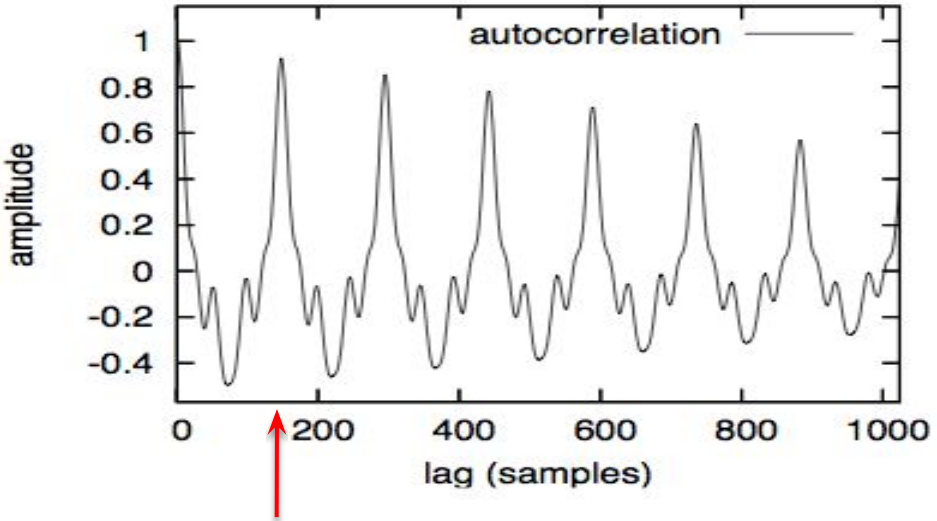
- Autocorrelation Function



# Time Domain: Auto Correlation (AC)



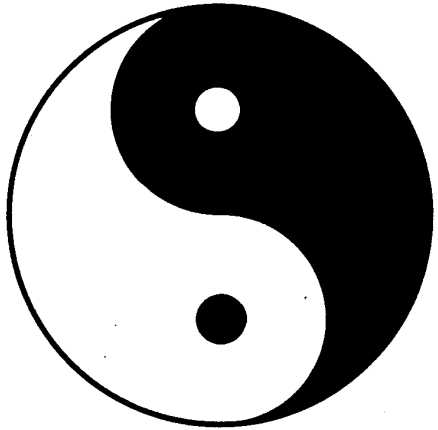




Frequency $f_0 = 1/\tau_0$

# Time Domain: YIN

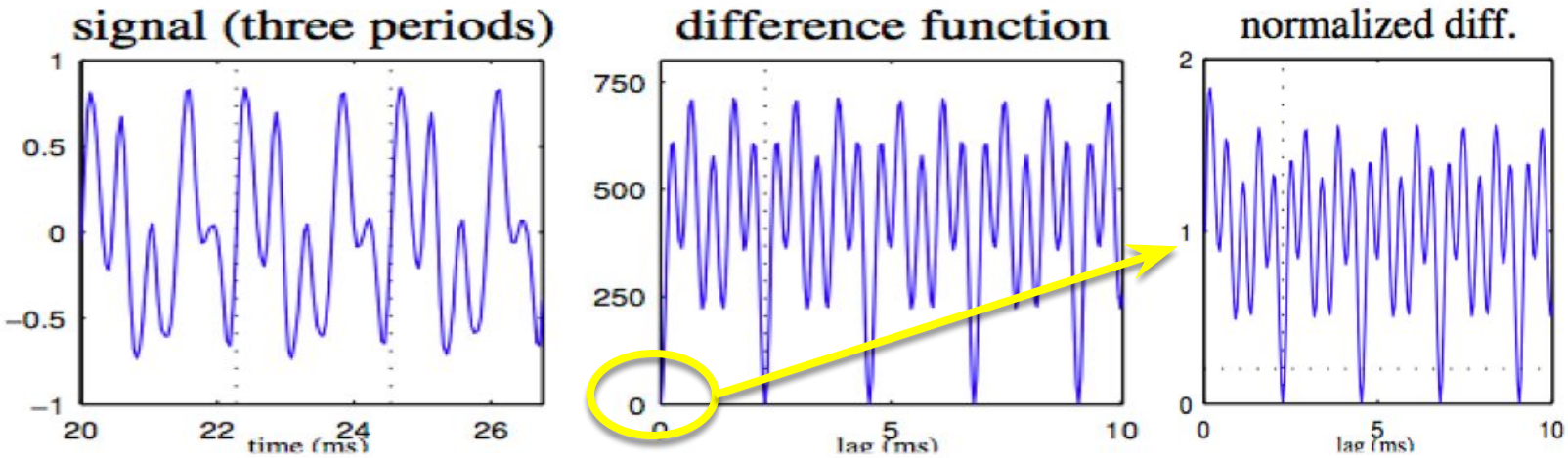
- Previous AC methods: look for maximum product
- YIN: look for minimum difference

$$d(\tau) = \sum_{n=0}^{N-1} (x(n) - x(n+\tau))^2$$



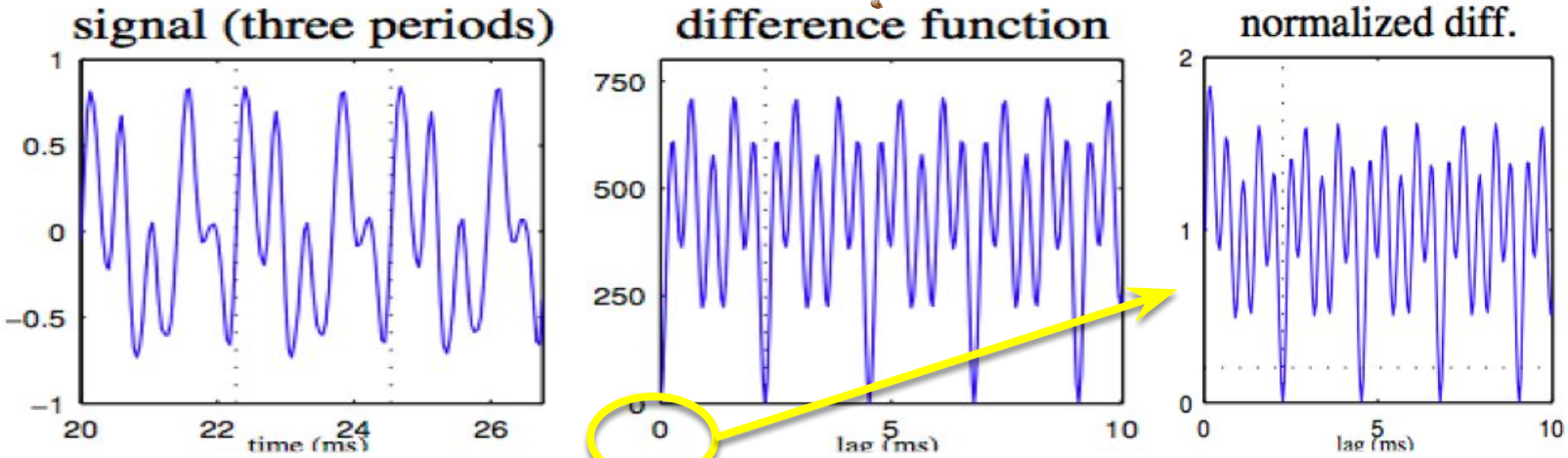
(de Cheveigne and Kawahara, 2002)

# Time Domain: YIN



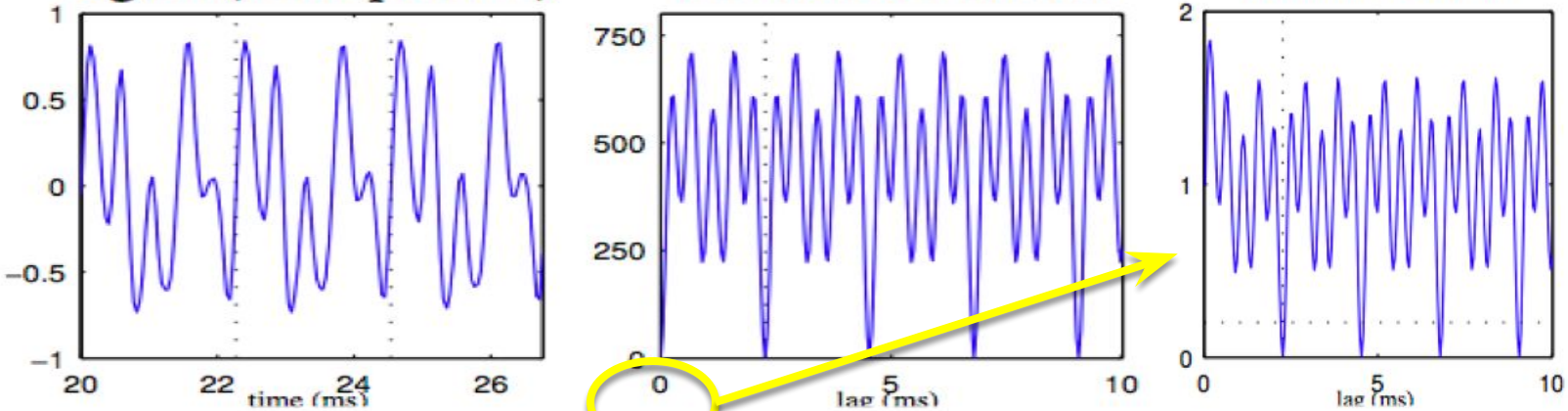

$$d(\tau) = \sum_{n=0}^{N-1} (x(n) - x(n+\tau))^2$$

# Time Domain: YIN



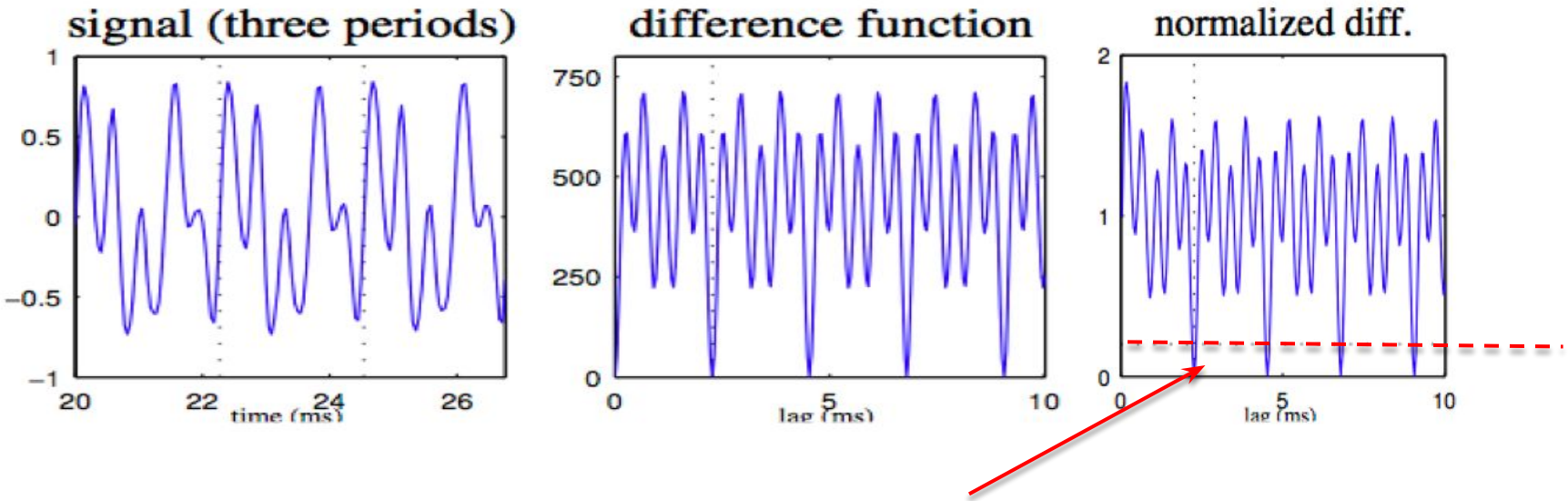

$$= \begin{cases} 1 & \tau = 0 \end{cases}$$

# Time Domain: YIN




$$= \begin{cases} 1 & \\ \text{  } \Big/ Average(\text{  }) & \end{cases} \quad \begin{matrix} \tau = 0 \\ otherwise \end{matrix}$$

# Time Domain: YIN




**Last step:** Select 1st minimum under the threshold → fast

# Time Domain: YIN

- Final Steps:
  - If signals are not perfectly harmonic, do parabolic interpolation (optional)
  - Find best local estimate (similar to median smoothing)
- YIN has the best accuracy in many of the comparisons I've read

# Frequency Domain

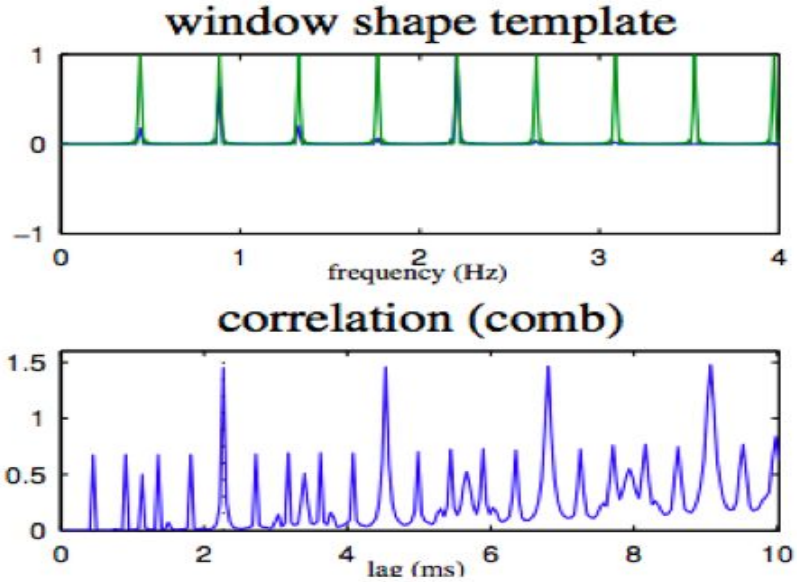
*Spectral Comb Filtering*

*Maximum Likelihood*

*Harmonic Product Spectrum*

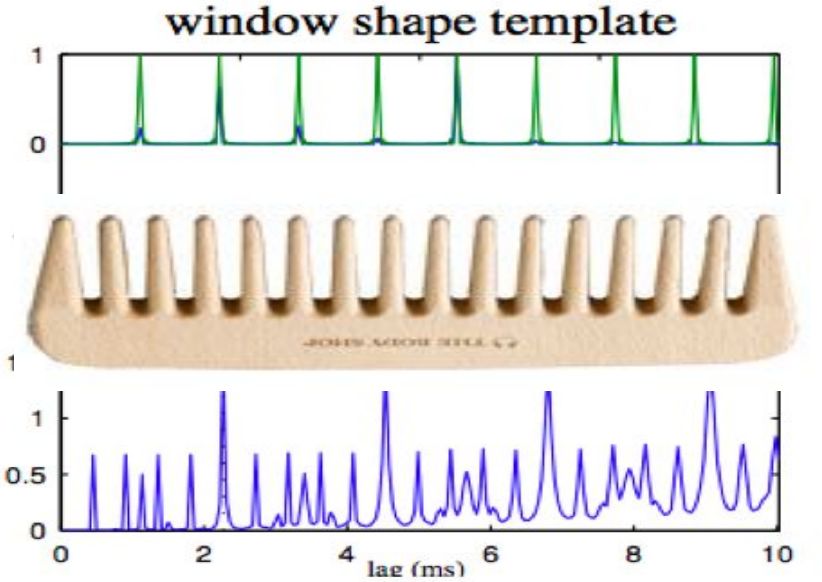
# Spectral Comb Filtering

- Use a comb-shaped band-pass filter at the f0 and each of the harmonics.



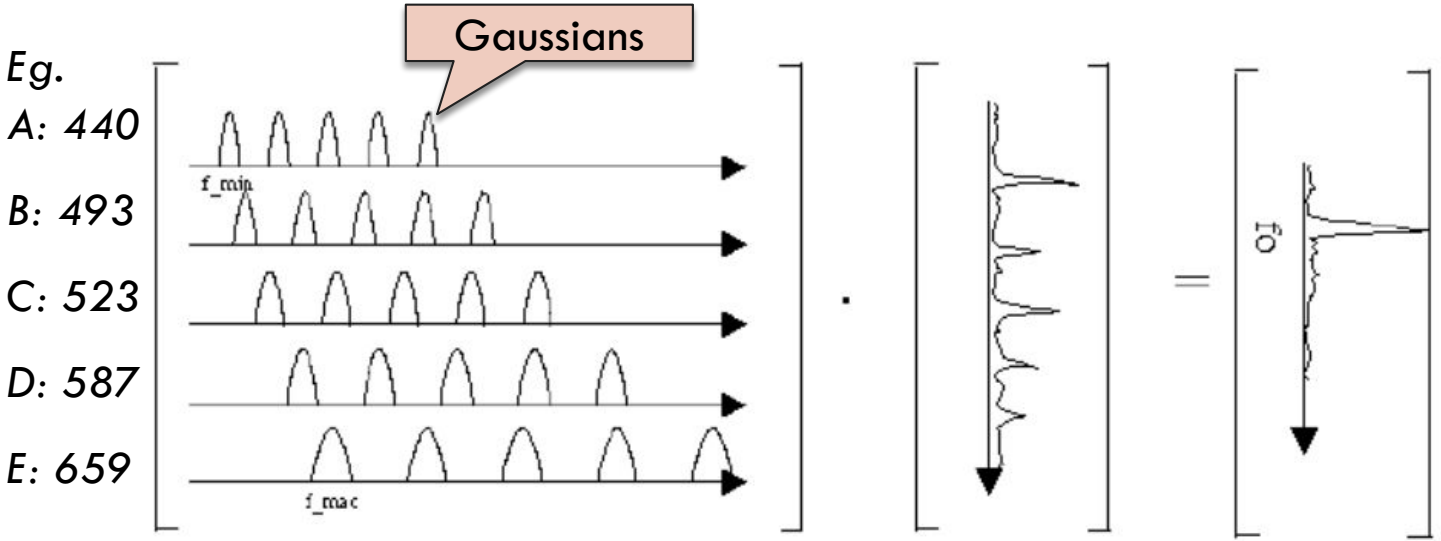
# Spectral Comb Filtering

- Use a comb-shaped band-pass filter at the f0 and each of the harmonics.



# Maximum Likelihood

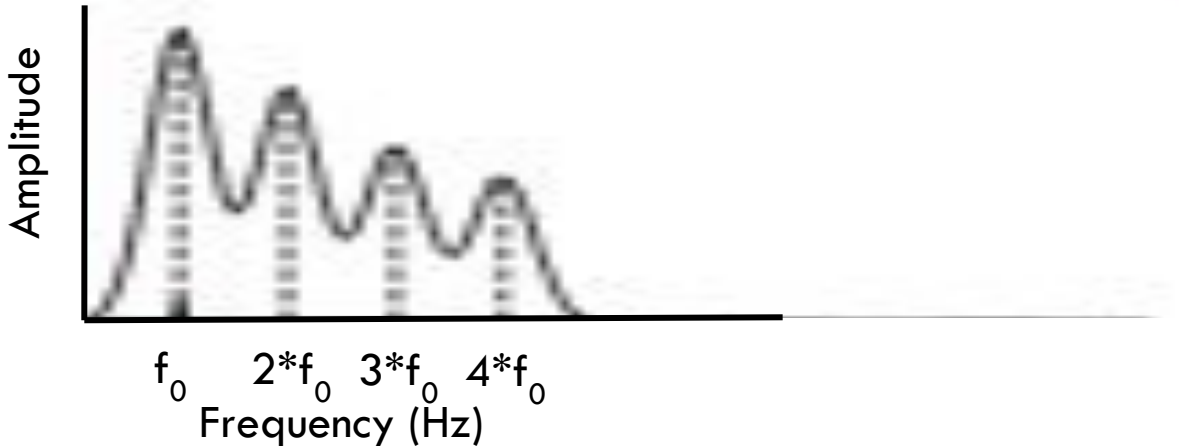
- Convolve the signal with a template for each note to be detected (eg. A=440Hz, B=493 Hz,…)



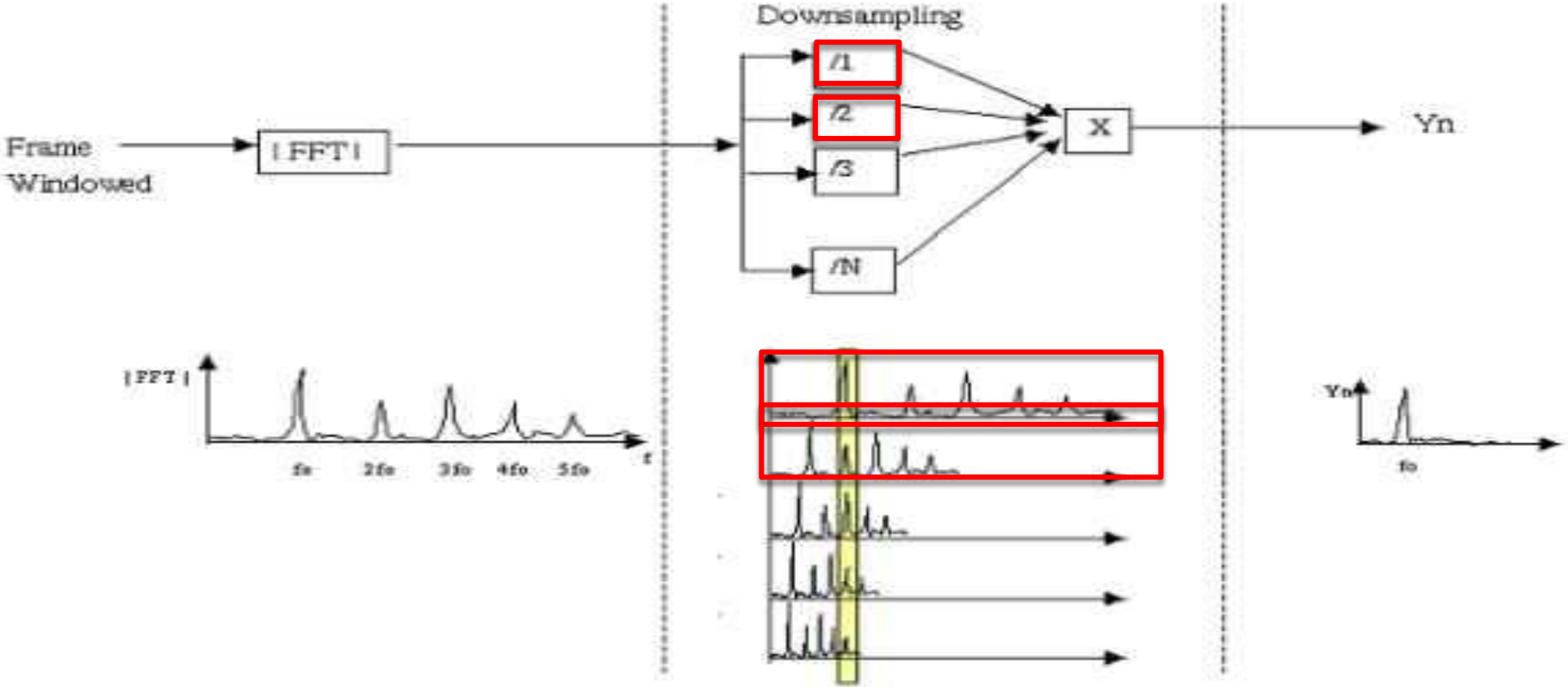
# Maximum Likelihood

- A kind of Pattern Matching
- **Good** for fixed pitch instruments (eg. piano)
- **Not** good for voice, violin, etc. because vibrato can produce in-between pitches not in our template database

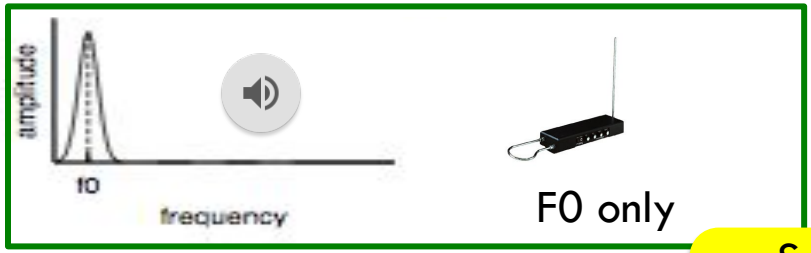
# Frequency Domain Reminder



# Harmonic Product Spectrum



# More than just pitch!



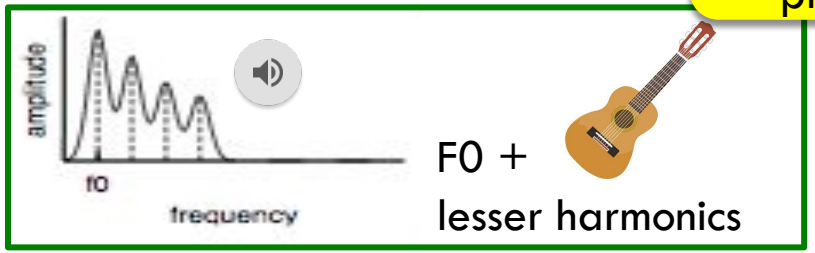



F0 only




F0 +
strong harmonics

Same perceived pitch!






F0 +
lesser harmonics




No F0 +
strong harmonics