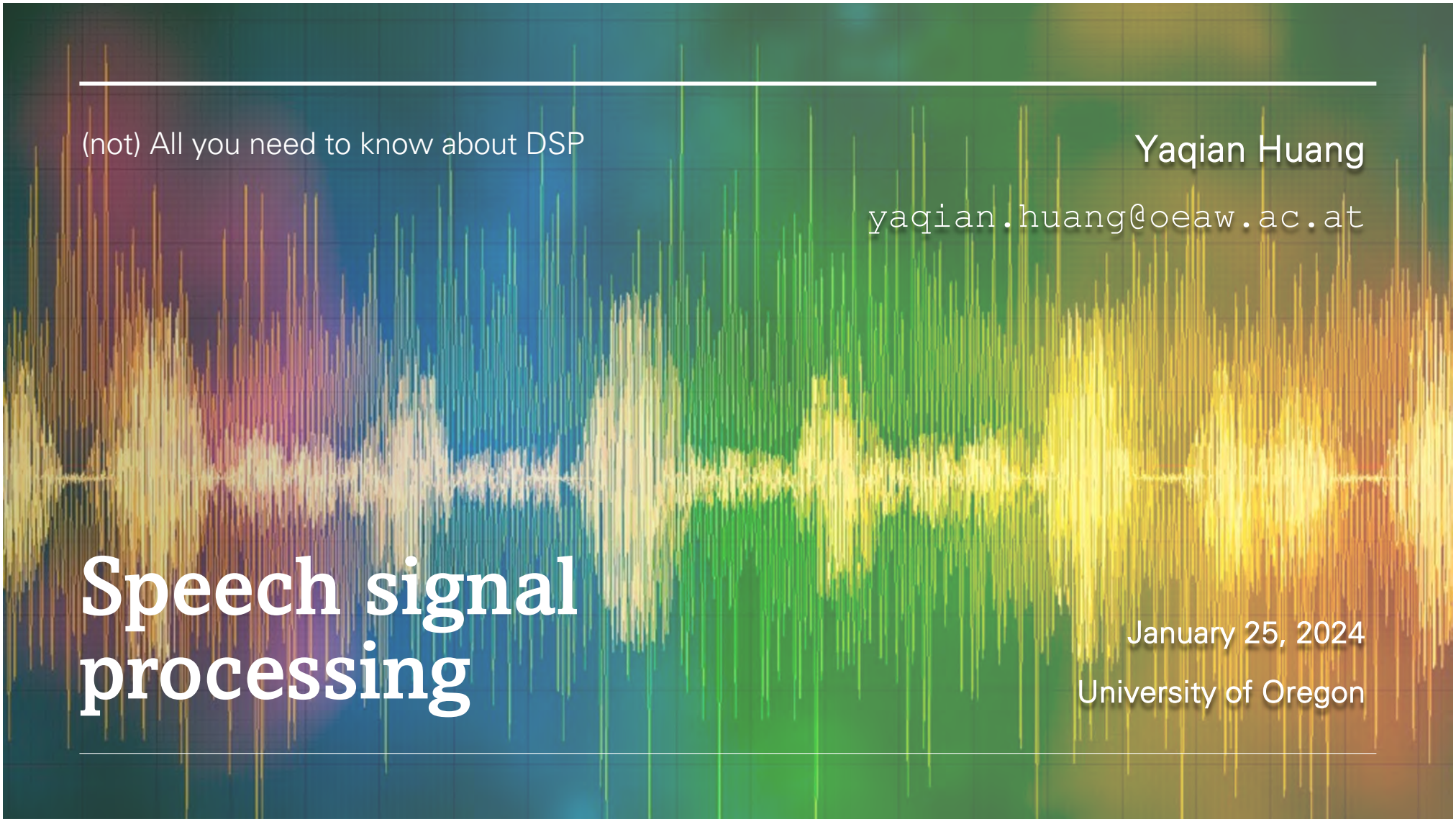


(not) All you need to know about DSP

Speech signal processing

Yaqian Huang

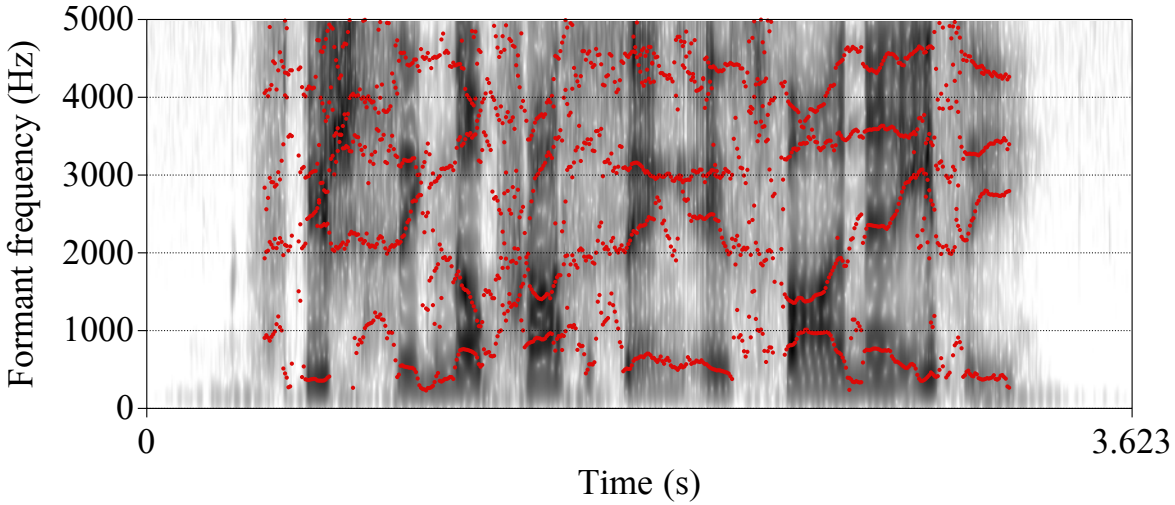
yaqian.huang@oeaw.ac.at

January 25, 2024

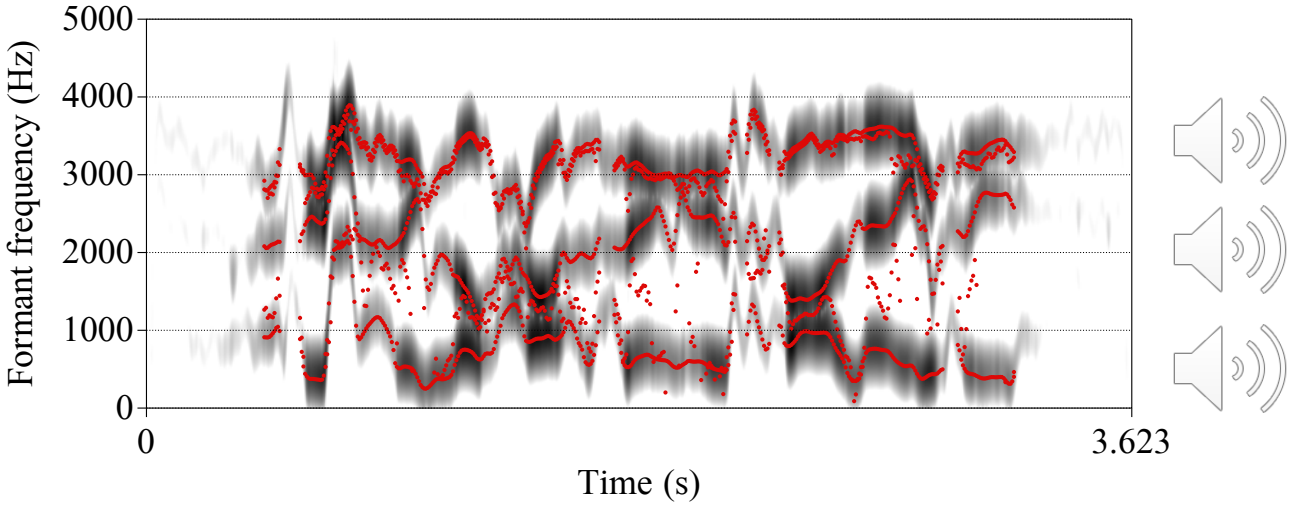
University of Oregon

Mysterious audio

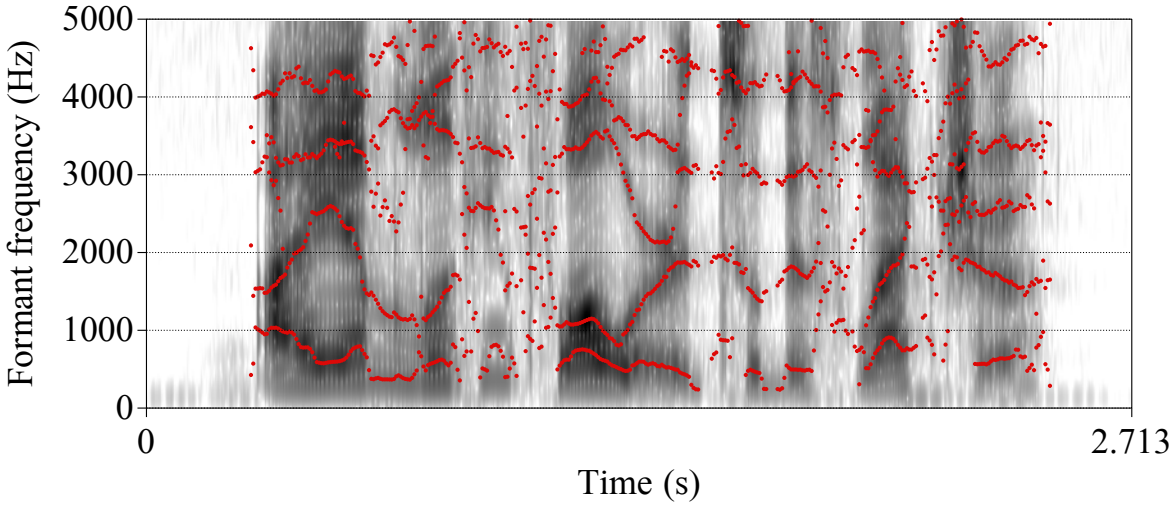
Original speech 1



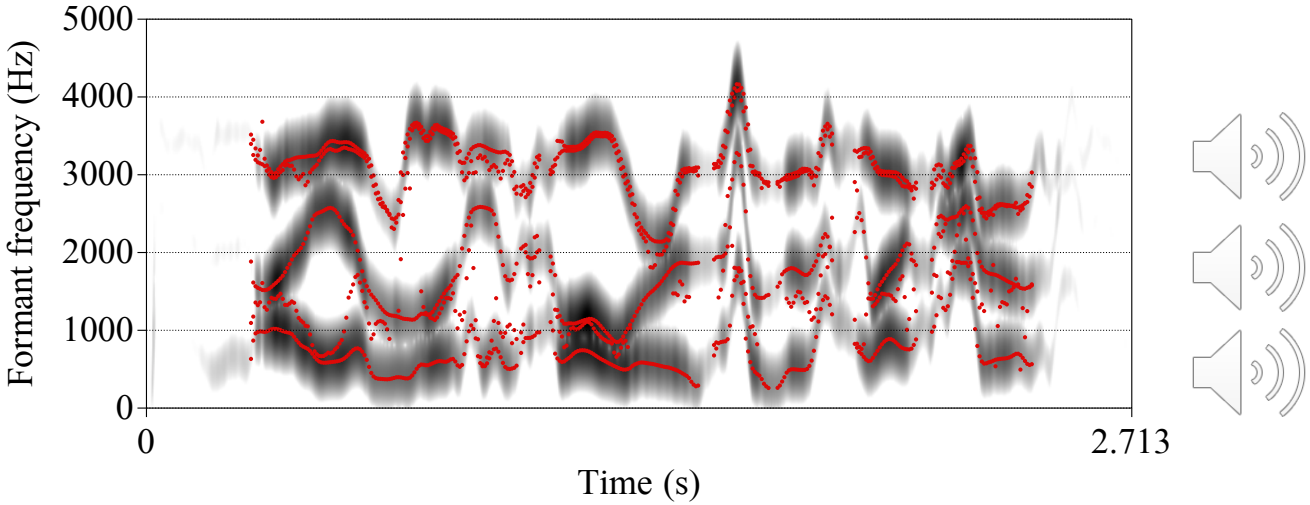
Sinewave speech 1



Original speech 2



Sinewave speech 2



Abstraction of speech signal

- Sinewave speech is a type of synthesis technique to model speech and test speech perception theories

- Our mental representation of speech
  - Distill to the most informative cues
  - May consider other factors, e.g., idiosyncratic, accents, etc.

- How much does a computer need?

→ Time to learn about digital signal processing

Goals of today's lecture

- Gain understanding of fundamentals of speech signal processing
- Be familiarized with techniques of pitch extraction in the time domain
- Practice speech processing & synthesis techniques in Praat though hands-on learning

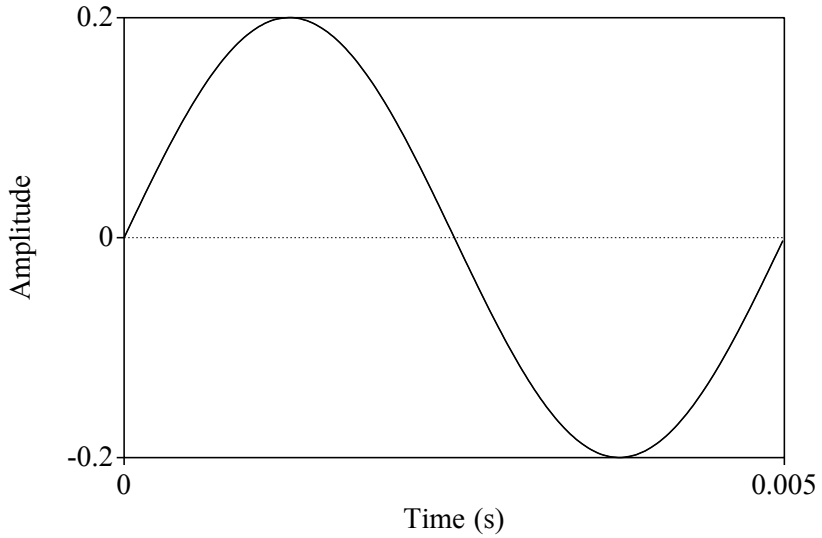
Outline

- What does DSP mean and what tools do we need
- Fundamentals: sampling, quantization, aliasing, filtering
- Pitch extraction (in time domain)
- Hands-on speech synthesis activities using PSOLA

Let computer understand speech

- Sound is analog, computer is digital

- We need an Analog-to-Digital converter
- Microphones

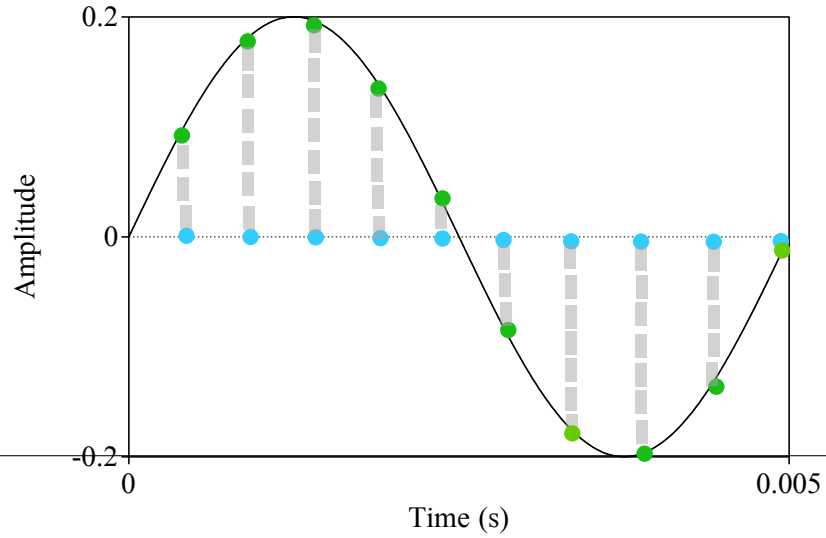


Analog-to-Digital conversion

- Two processes are involved:

Sampling is the process of making the x-axis (time) discrete

Quantization is the process of making the y-axis (amplitude) discrete



Quantization: how much do we sample?

- Sound waves create air pressure at *real* numbers
- A computer can only store a *finite* precision of such numbers
- Discretize every *real* number many times per second to obtain a *finite* precision: rounding, truncation of numbers

- The precision is defined within a range of bits (binary 0/1)
  - 16 bits per sample are commonly used

Sampling: how often do we sample?

- Sampling frequency or sampling rate (Fs): N times a second, measured in samples per second (Hz)

Example: A 1 Hz sampling rate means one sample per second

High sampling rates mean better signal quality!

- How precise do we need? Depends on …
  - How much we can hear → the upper limit
  - How much is required to store the original frequency → the lower limit

How much do we use and hear?

- Vocal pitch: 75 – 500 Hz (setting in Praat)
- Formants: 250 – 3000 Hz
- Fricatives: > 3000 Hz

- We can hear up to 20,000 Hz
  - Which degrades with age → high frequency goes away first

- Why is CD sampling rate 44,100 Hz?

Let's find out by listening …

Praat command:
Convert – Resample…

At what sampling rate do you start to think it's bad quality or hear something else?

44,100 Hz (Nyquist frequency = 22,050 Hz)

22,050 Hz 

11,025 Hz 

6000 Hz 

3000 Hz 

1500 Hz 

800 Hz 

Nyquist Theorem

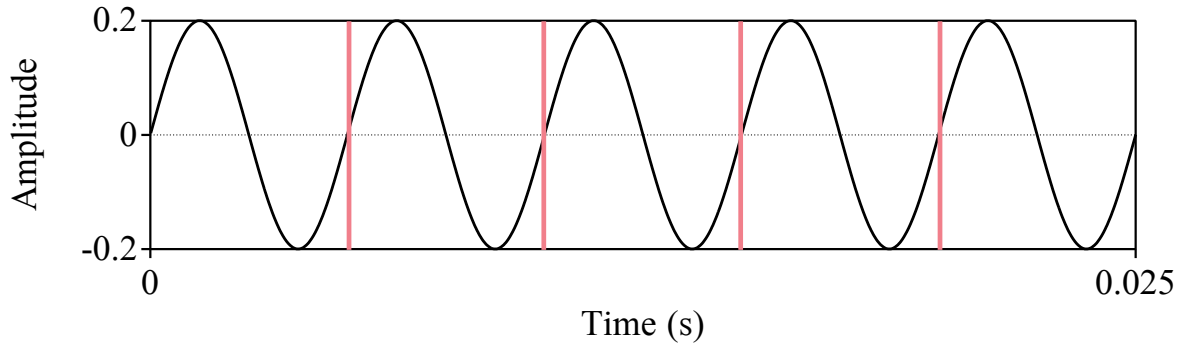
- The highest frequency that can be **unambiguously** captured by a sample signal is **one half the sampling rate**: Fs/2

➢ to store a frequency, we need a sampling frequency **at least 2 times the signal frequency**

- How many samples a second is required
  - To capture an 8000 Hz signal? 16,000 Hz?

Aliasing

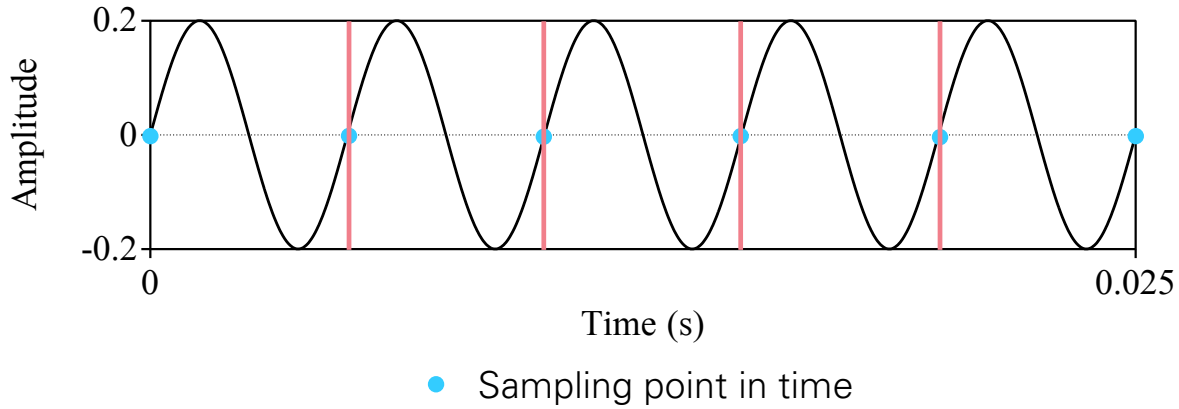
- Causes different signals to become indistinguishable ("aliases" of one another)
- We will try sampling a signal using different frequencies
- An example of a 200 Hz tone: 5 cycles



Aliasing

- Let's take 1 sample per cycle from this 200 Hz tone

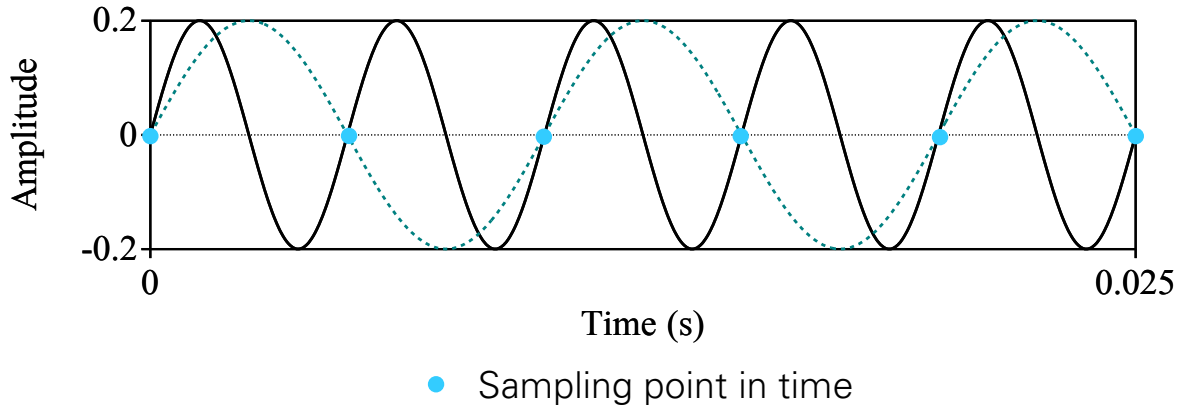
- The sampling frequency = 200 Hz, the same as f0



Aliasing

- Causes different signals to become indistinguishable

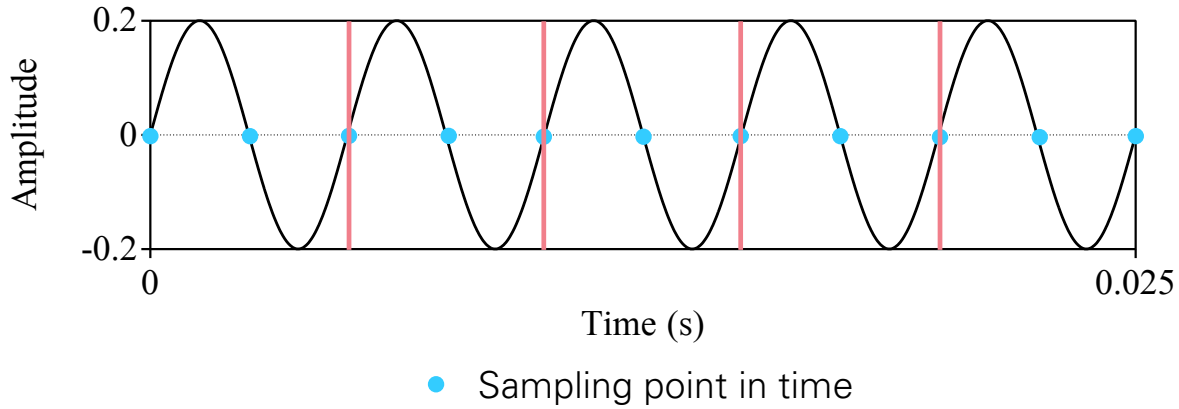
- At the sampling frequency of 200 Hz, we cannot tell the 200 Hz tone from a 100 Hz tone



Nyquist Theorem wins

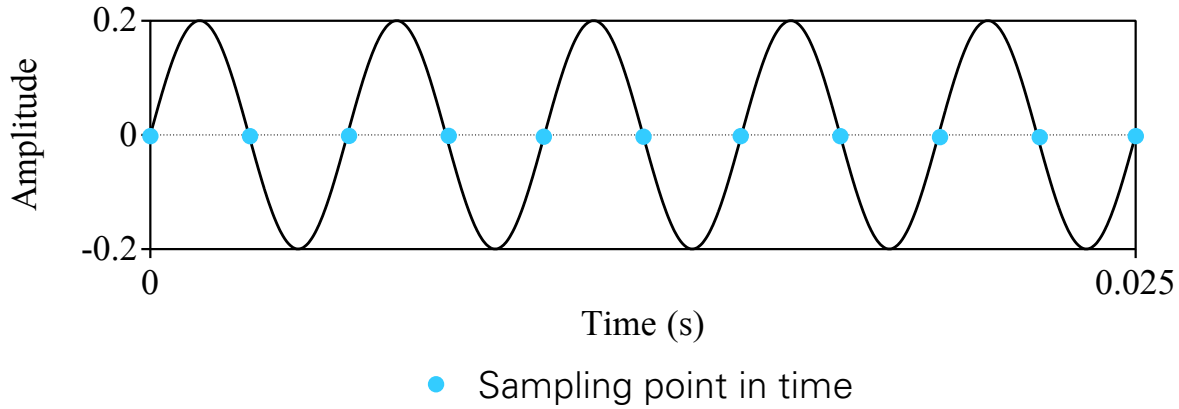
- Let's try again: double our samples in time for the 200 Hz tone

- We increased the frequency of taking samples from the signal: 2 samples/cycle



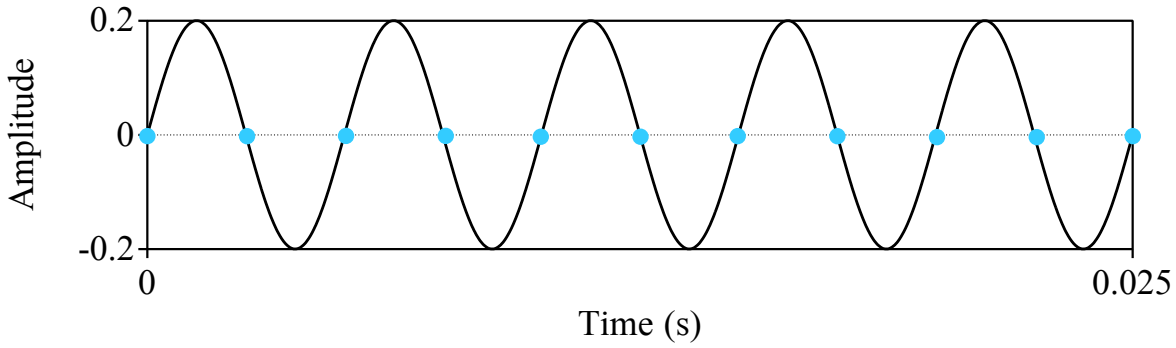
Nyquist Theorem wins

- Let's try again: double our samples in time for the 200 Hz tone
- We increased the frequency of taking the sample from the signal
- What is the sampling frequency now?



Nyquist Theorem wins

- Let's try again: double our samples in time for the 200 Hz tone
- We increased the frequency of taking the sample from the signal
- Now the sampling frequency = 400 Hz = 2 x 200 Hz **Two times of the f0!**



Sampling frequency examples

- Telephone standard: 8 kHz (narrow band)
- Most recordings do well with 16 kHz (wide band) → Most speech features fall below 8 kHz

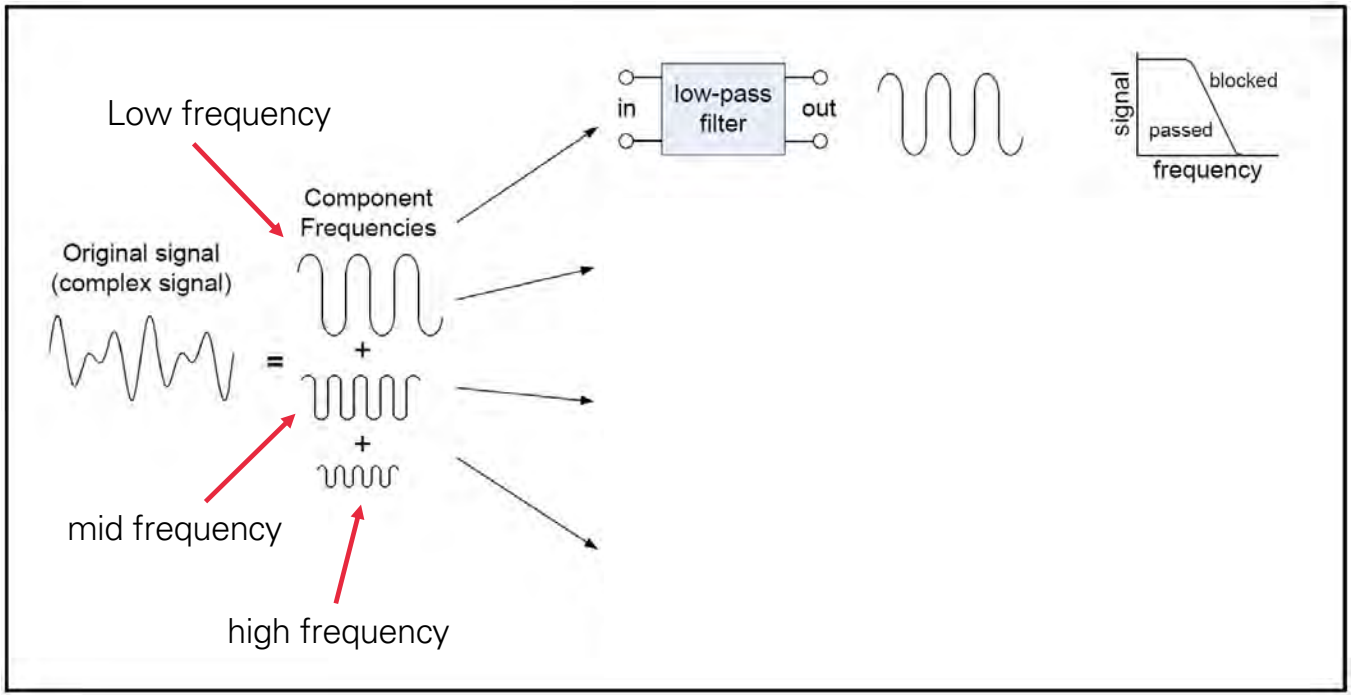
- CDs are at 44,100 Hz
- DVDs are at 48,000 Hz
- High-End Audio DVDs are at 96,000 Hz

- Some people want 192,000 Hz, and super audio CDs are even at 2,822,400 Hz!
- Likely they are dolphins → Why? Dolphins hear 2-200,000 Hz

Filtering

- Relative modification of amplitudes of different frequencies

- Low-pass filters
- High-pass filters
- Band-pass filters
- Notch filters



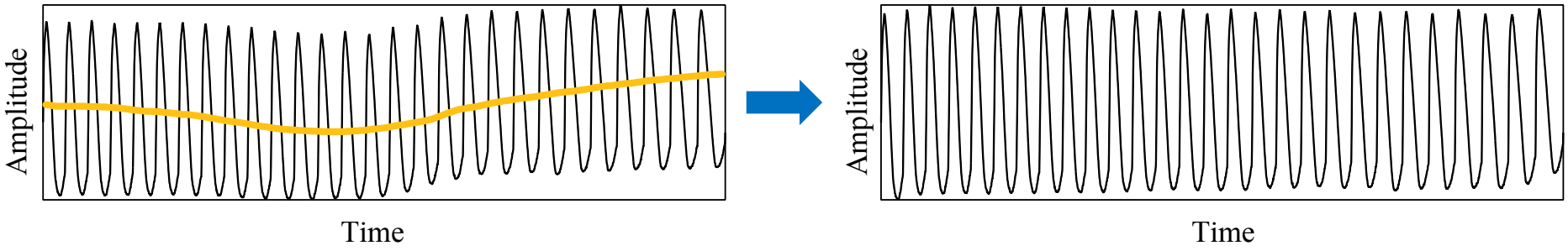
- When would you use which?

Example of filtering

Praat command:
Filter (Pass Hann band)…

- We may often encounter signals with noise: low-frequency buzz, high-frequency noise

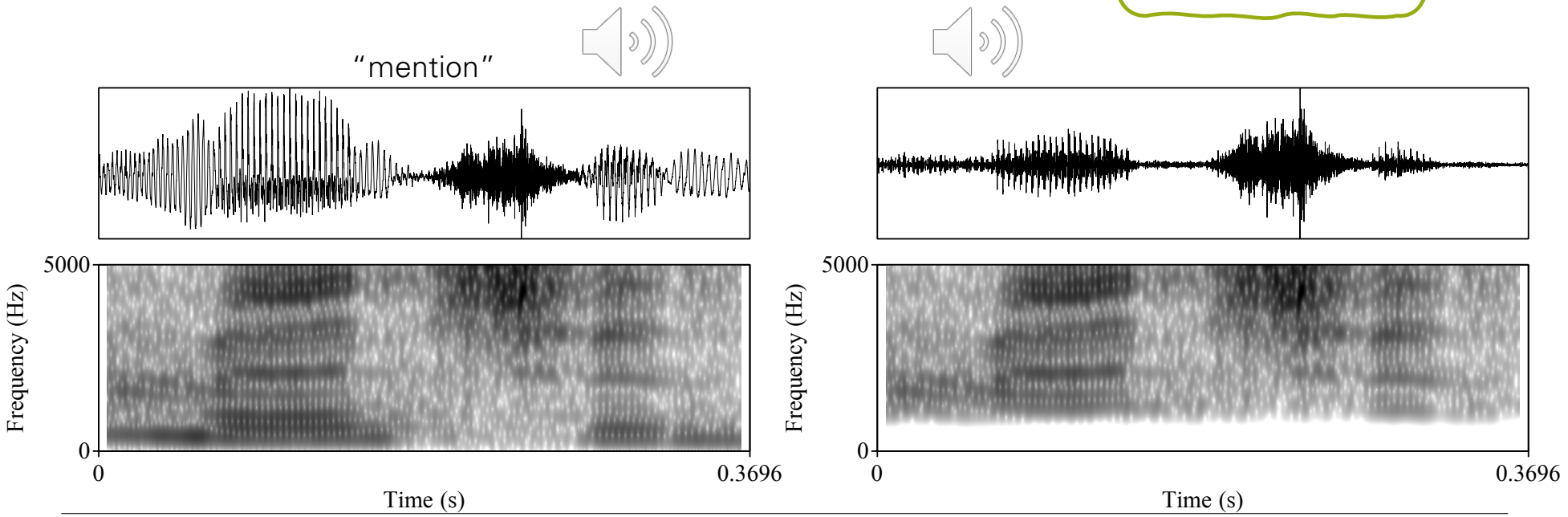
- Remove low frequency buzz around 40 Hz: electroglottographic (EGG) waves



Scenario checking 1

- Which filter is used here?

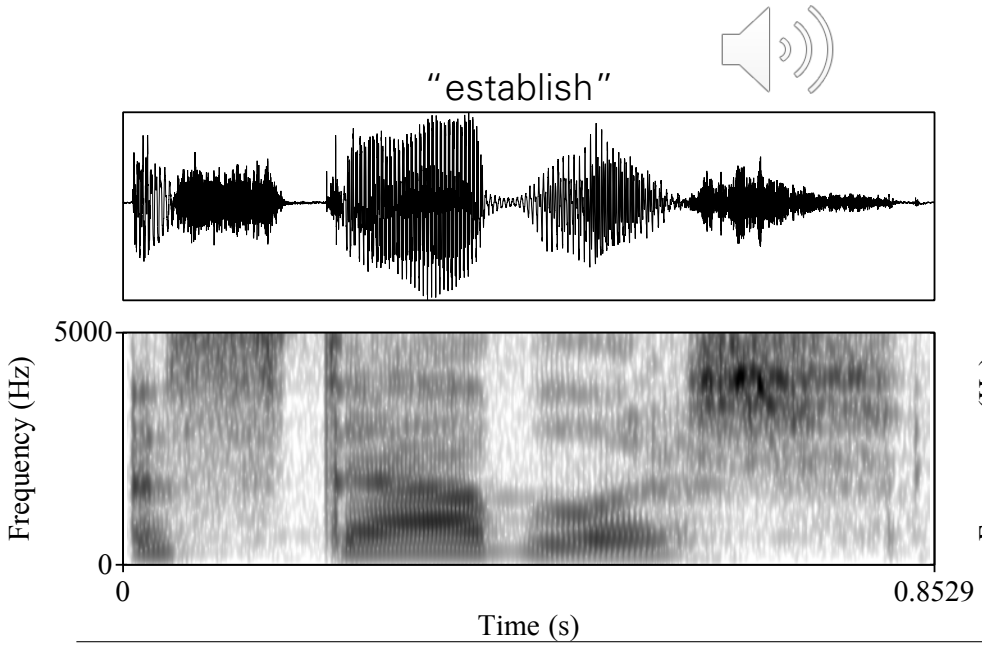
- Low-pass filters
- High-pass filters
- Band-pass filters
- Notch filters

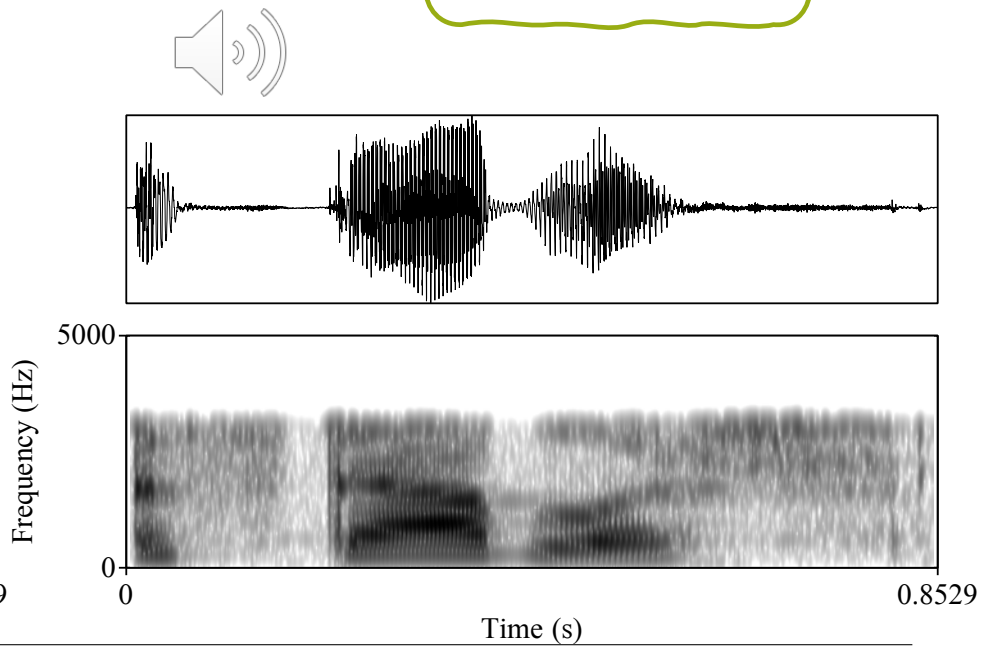


Scenario checking 2

- Which filter is used here?

- Low-pass filters
- High-pass filters
- Band-pass filters
- Notch filters

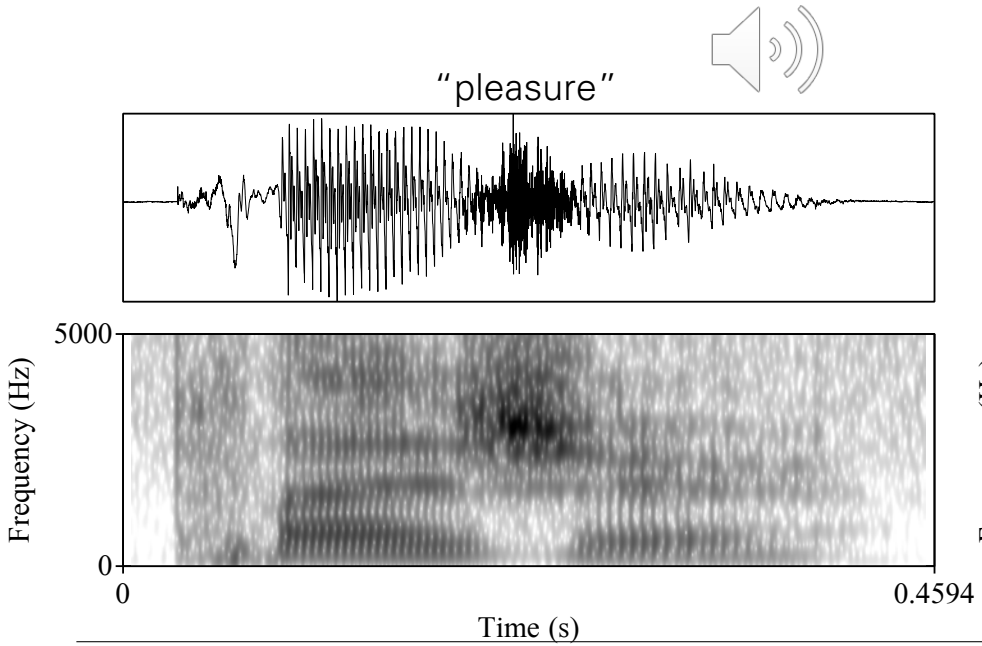


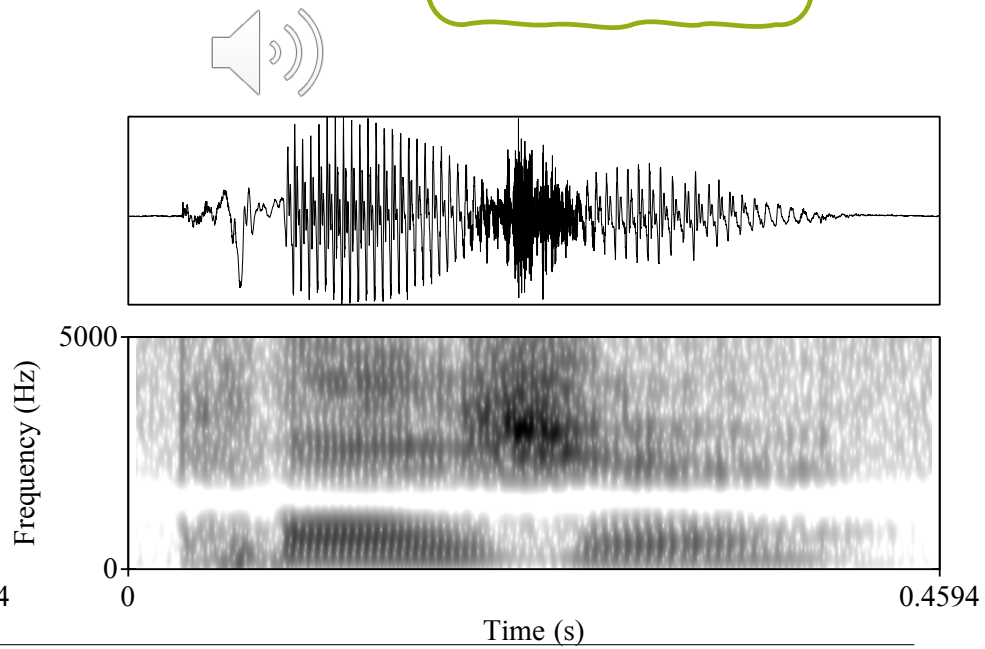


Scenario checking 3

- Which filter is used here?

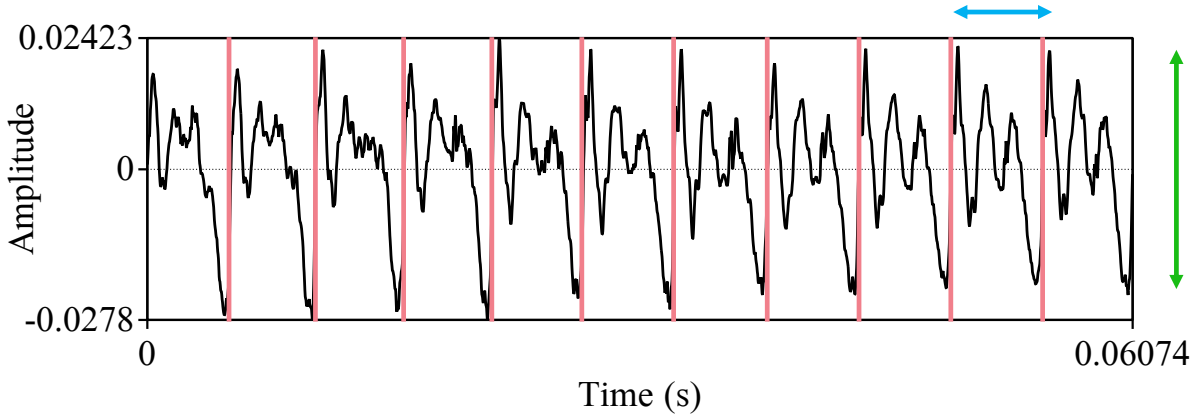
- Low-pass filters
- High-pass filters
- Band-pass filters
- Notch filters





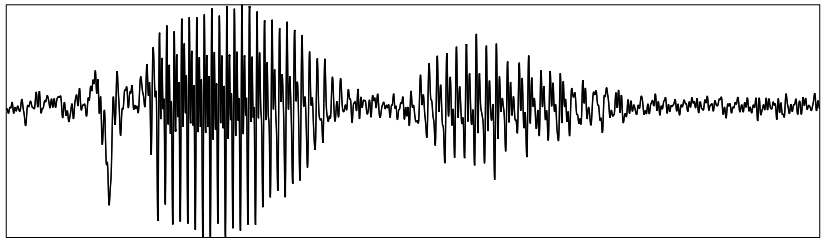
How to determine pitch (in the time domain)?

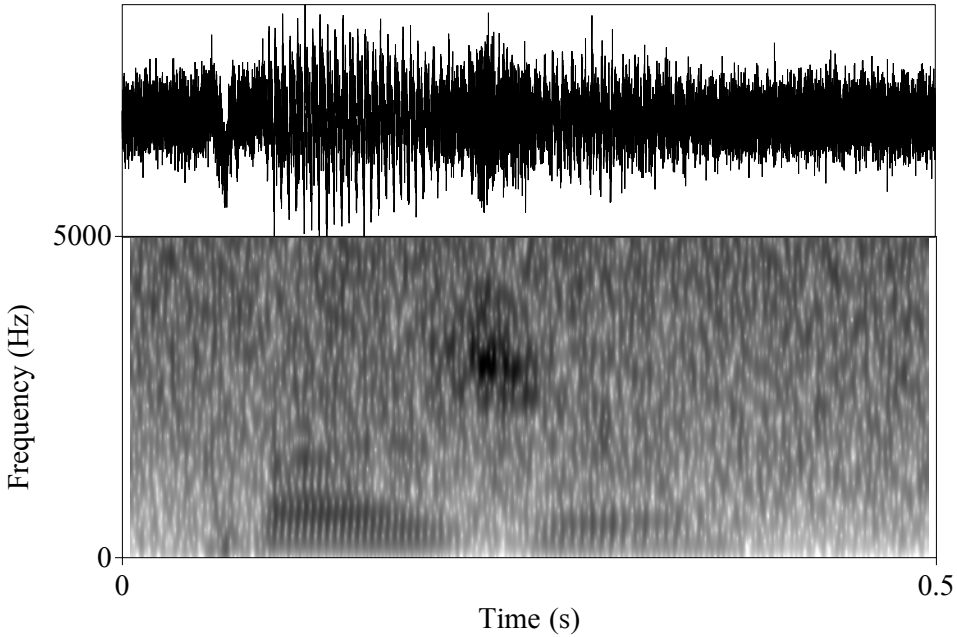
- We look for the period of a repeating cycle in voiced speech
- Sine waves are easy because of clearer time periods
- Speech is quasi-periodic!



Challenges to pitch extraction

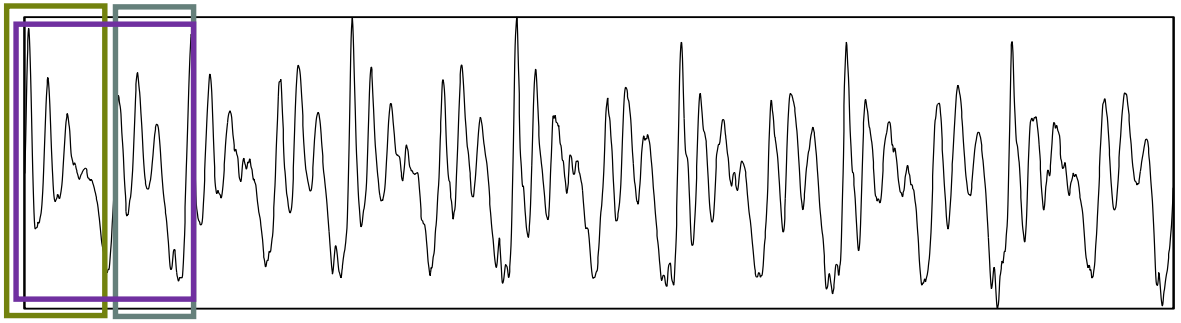
- Noise
  - Pitch is usually < 600 – 700 Hz
  - We can use low-pass filter





Challenges to pitch extraction

- Multiple f0s
  - e.g., problem with period doubling →
    - You will hear more in my talk!

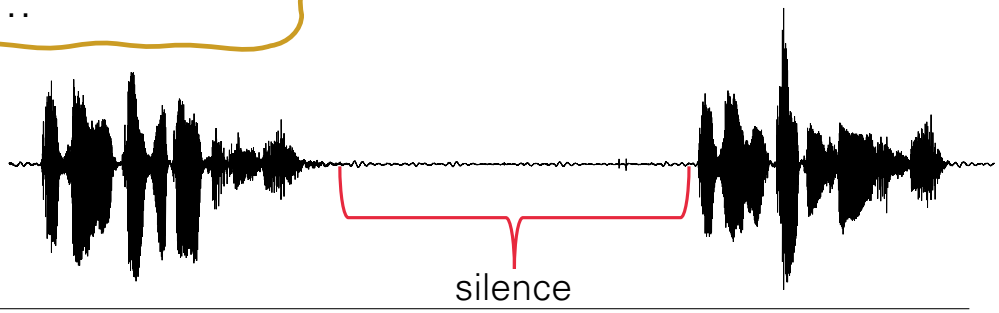


  - Multiple talkers present → speaker diarization

Challenges to pitch extraction

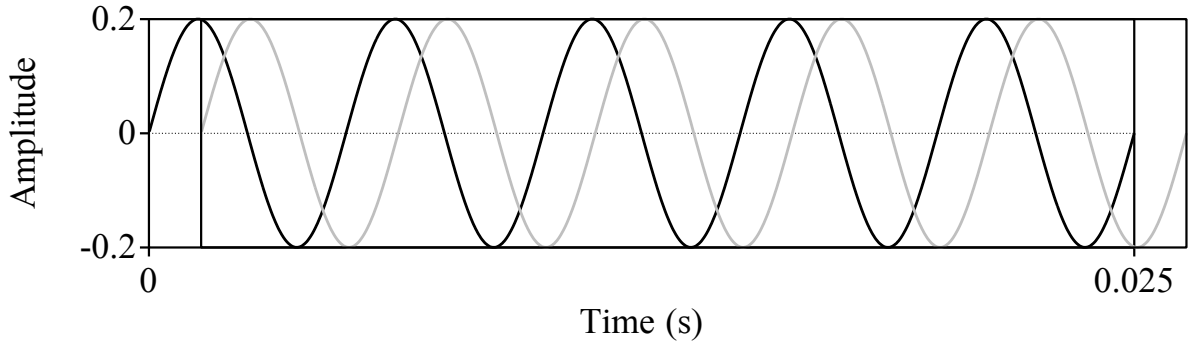
- Voice activity detection
  - Presence of silence, voiceless sounds
  - We can use energy difference, spectral properties

Praat command:
- Annotate – To TextGrid (silences)…
- Annotate – To TextGrid (voice activity)…



Autocorrelation

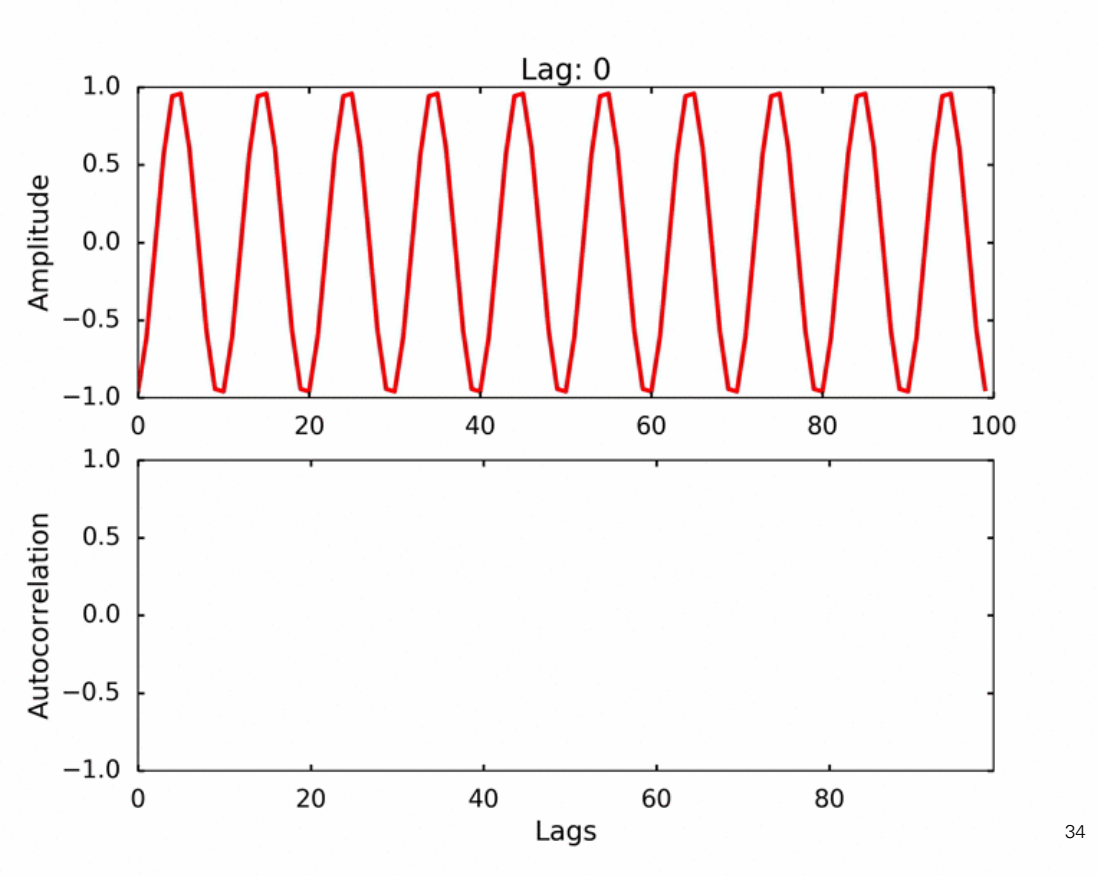
- We extract pitch by determining the time difference between two signals where one is a delayed version of the other



- When the time difference = a period of cycle, two signals will overlap for the first time
- How do we know two signals overlap?

Autocorrelation

- Sum the multiplication of each time sample from two signals at a certain lag: samp1_t1*samp2_t1+ samp1_t2*samp2_t2+ …
- We observe that when lag = 10, the autocorrelation is second to the largest
- At the integer multiples of 10, autocorrelation is always at peak
- This lag = 10 should be the pitch period



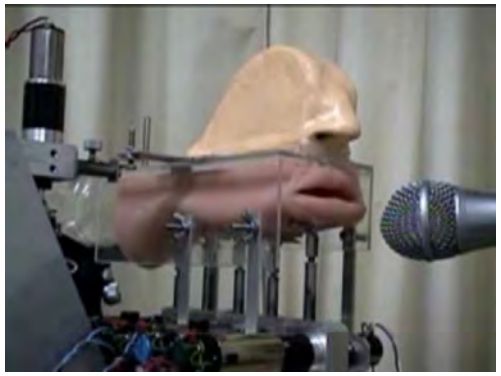
Hands-on: pitch modification in PSOLA

- Pitch-Synchronous-Overlap-Add
- Copy synthesizer → provides control of both F0 and duration
  - Often used in music generation, text-to-speech synthesis

Praat command:
- Read in the file named "sentence-psola.wav"
- Manipulate - To Manipulation…
- Stylize pitch…
- Try removing & adding pitch points in the middle pitch manip area
- Play (overlap-add)

Other synthesis techniques for research experiments

- Formant synthesis
  - uses a certain number of formants to represent sound
- Formula synthesis
  - E.g., Speech in noise
- Articulatory synthesis
  - Math model of vocal tract that can generate realistic sounds



https://www.youtube.com/watch?v=qobhDJ_vEOc

Speech processing & synthesis applications

- Automatic speech recognition
  - Acoustic modeling
  - Language modeling

- Text-to-speech synthesis
  - Concatenative synthesis (uses PSOLA)
  - Statistical parametric synthesis

- Generative AI
  - Neural network

https://www.youtube.com/watch?v=qizNQKzatXA

Concatenative synthesis

- Prep: Have a database of real human recordings, divide them into segments which encode important phonetic information

← like the starters of speech production studies, right?

- Process:
  - Text modeling
  - Choose optimal chunks based on the goal (can be automatized)
  - Concatenate
  - Modify prosody & duration (uses PSOLA)

Wrapping up

- Speech signal processing is fun and easy
- Foundations: sampling, quantization, aliasing, filtering
- Time domain pitch extraction based on autocorrelation
- We are using DSP to model speech essentially!


(not) All you need to know about DSP
Yaqian Huang
yaqian.huang@oeaw.ac.at
Thank you!
January 25, 2024
University of Oregon

Optional: voice quality synthesis in KlattGrid

- Formant synthesizer → uses a certain number of formants to represent sound

Praat command:
- Open → Read from file… "va.wav"
- Praat → Open Praat script… "script_klatt_vq"
- In Script window: Run → Run

	Modal	Creaky	Breathy
Breathiness Amp	40	40	> 40
Open Phase	0.5	0.1 - 0.5	0.5 - 1
Spectral tilt	10	< 10	> 10
Flutter	0	0 - 1	0
Double Pulsing	0	0 - 1	0

↑ **Reference values**