Speech Recognition (DSAI 456)

Lecture 2

Mohamed Ghalwash
mghalwash@zewailcity.edu.eg

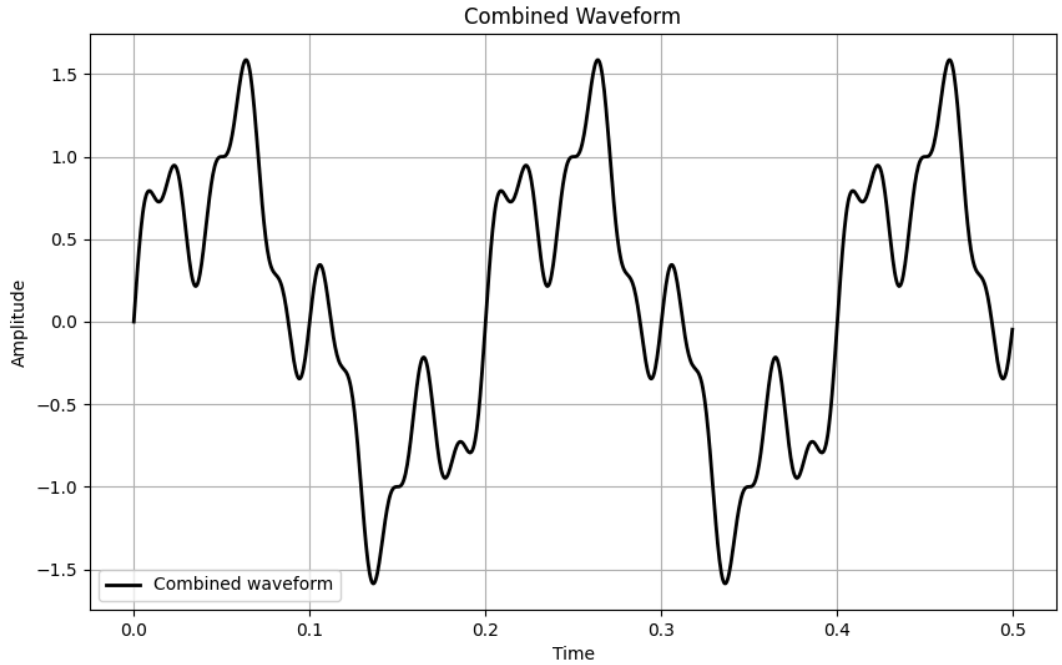
Lecture 1 Recap

- Phonetics, Syllables
- Signal Representation
- Analog to digital
- Transform the input waveform into a sequence of acoustic features
- Features: pitch, loudness, intensity, F0 tracks

What is

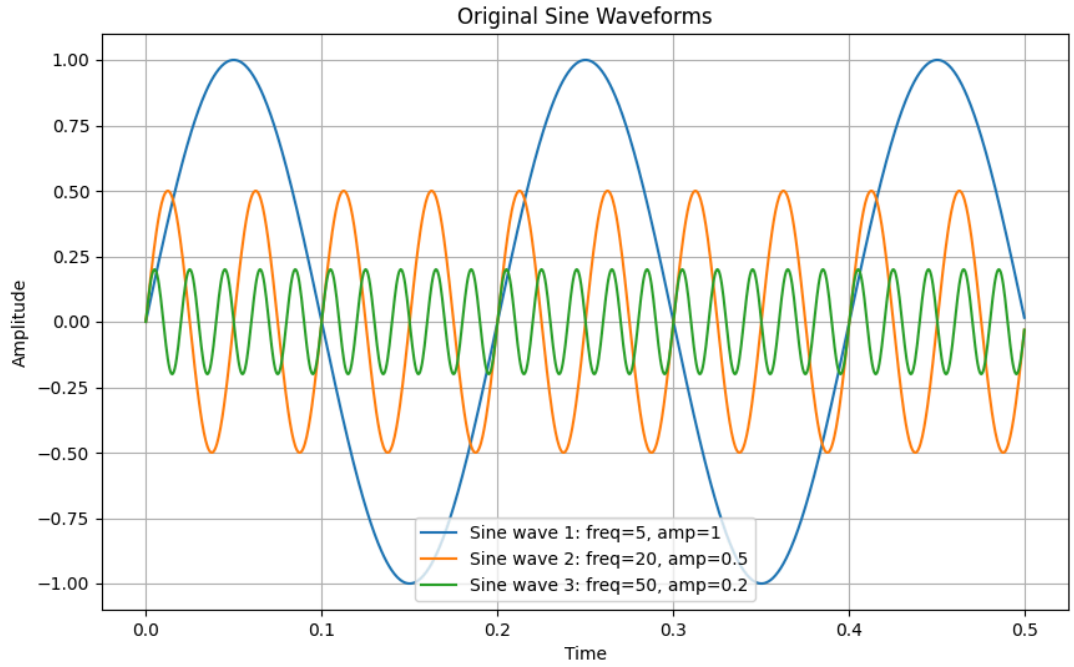
- Frequency Spectrum
- Frequency Bands
- Frequency Spectrogram
- Mel Spectrum

Frequency Spectrum



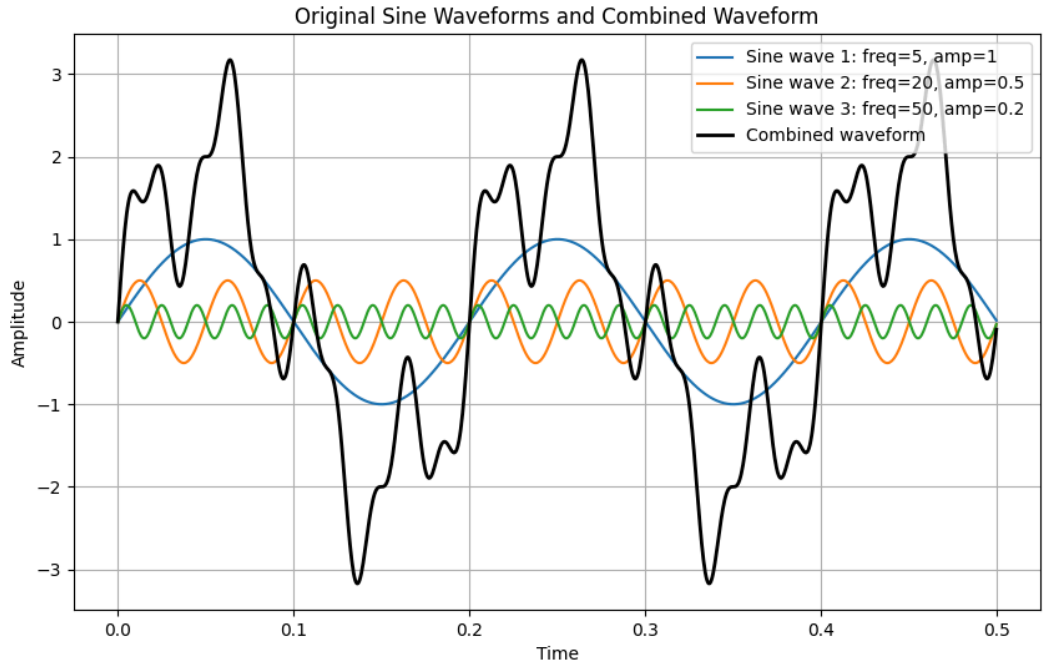
☹ It does not look like a sine waveform

Frequency Spectrum



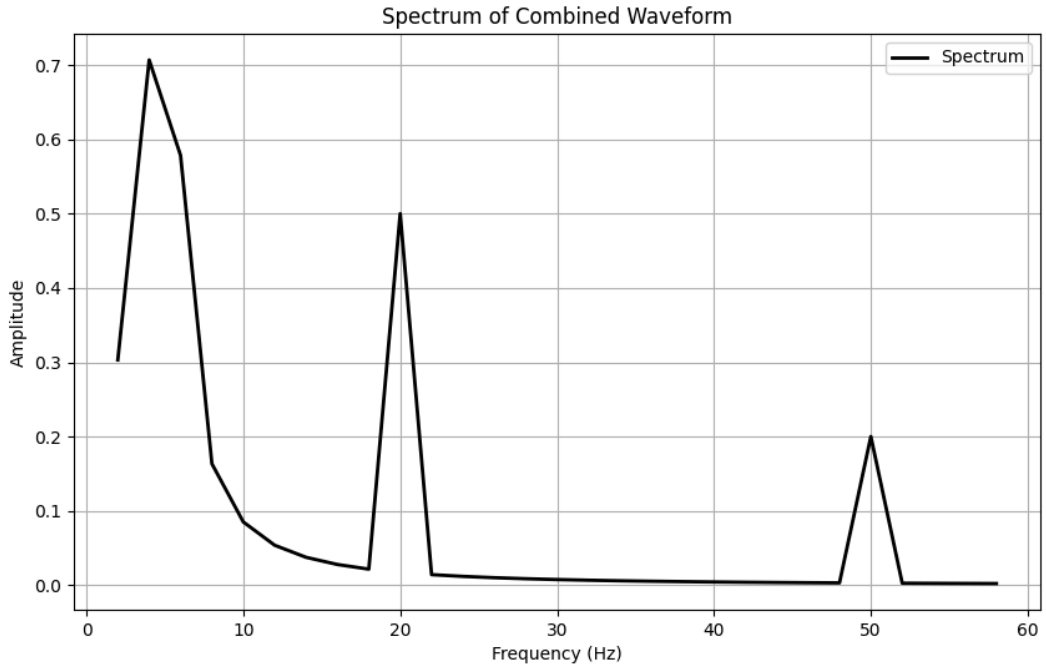
Maybe a combination of sine waveforms *with different frequencies and amplitude*

Frequency Spectrum

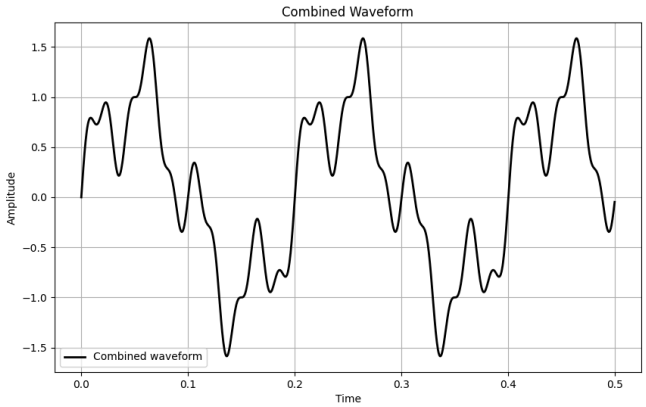


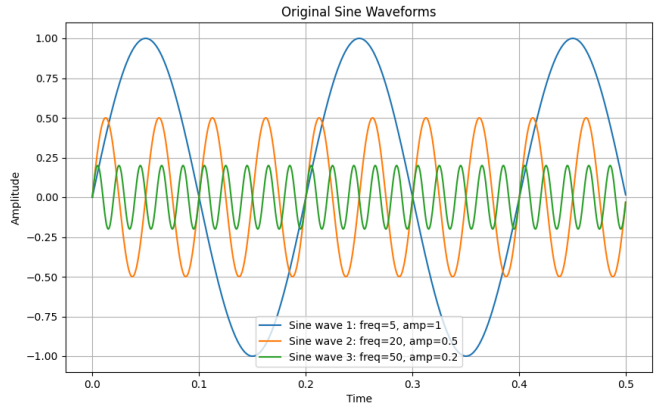
Aha! I got it

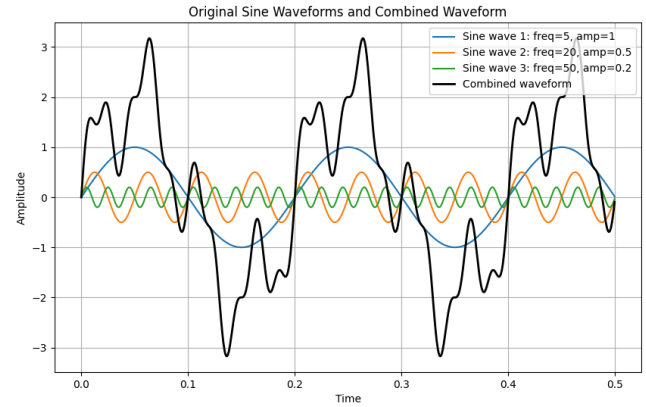
Frequency Spectrum

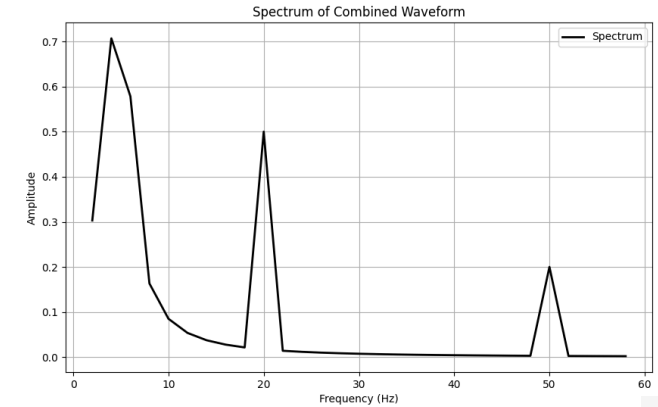


Done using Discrete Fourier Transform


Combined Waveform
Amplitude
Time
Combined waveform

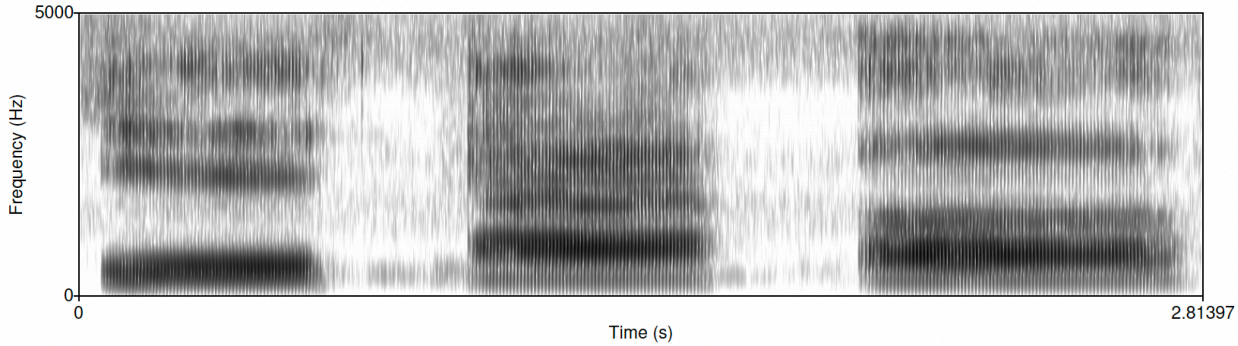

Original Sine Waveforms
Amplitude
Time
Sine wave 1: freq=5, amp=1
Sine wave 2: freq=20, amp=0.5
Sine wave 3: freq=50, amp=0.2


Original Sine Waveforms and Combined Waveform
Amplitude
Time
Sine wave 1: freq=5, amp=1
Sine wave 2: freq=20, amp=0.5
Sine wave 3: freq=50, amp=0.2
Combined waveform

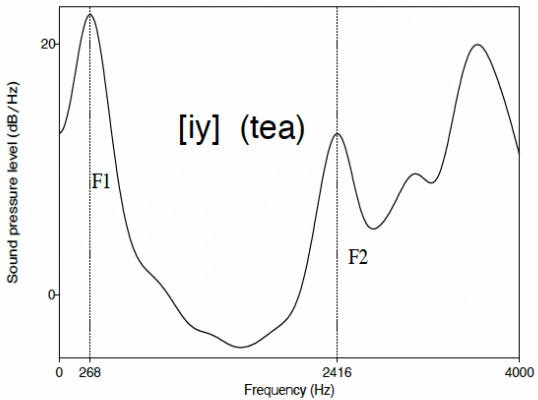

Spectrum of Combined Waveform
Amplitude
Frequency (Hz)
Spectrum

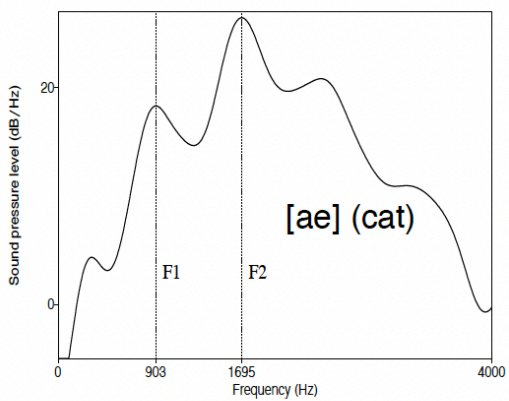
Frequency Bands

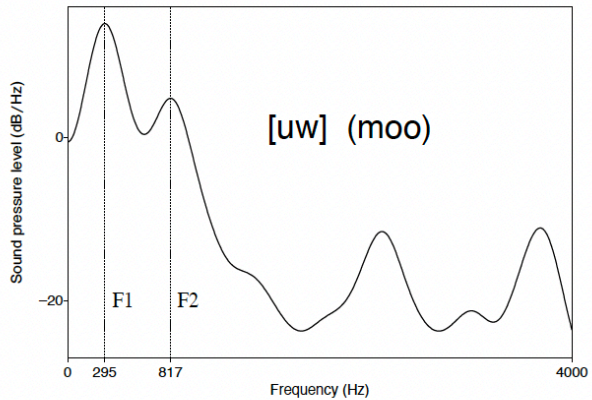
- Phones have characteristic spectral *signatures (Spectral peaks)*
- Inner ear computes the spectrum of the incoming waveform
- Spectrogram for three vowels

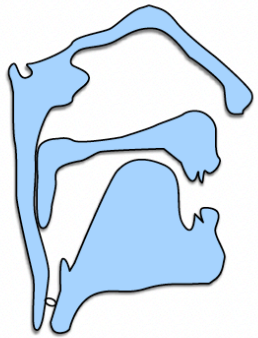


- Each dark bar is called a **formant** which is a frequency band that is particularly amplified by the vocal tract

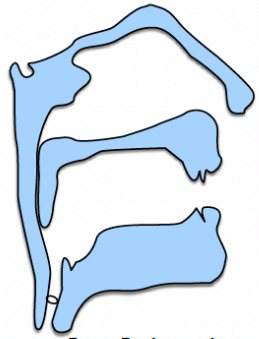




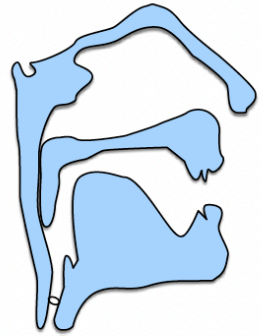




[iy] (tea)



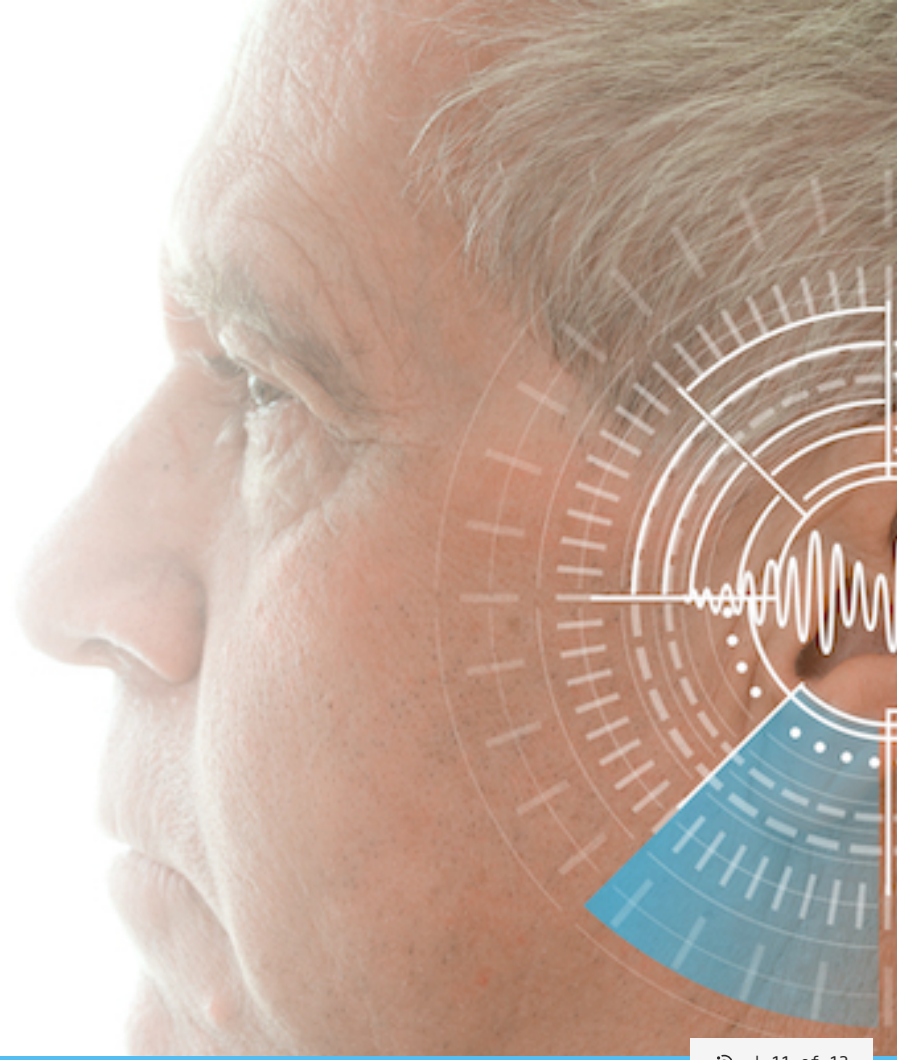
[ae] (cat)



[uw] (moo)

Issues with Hz

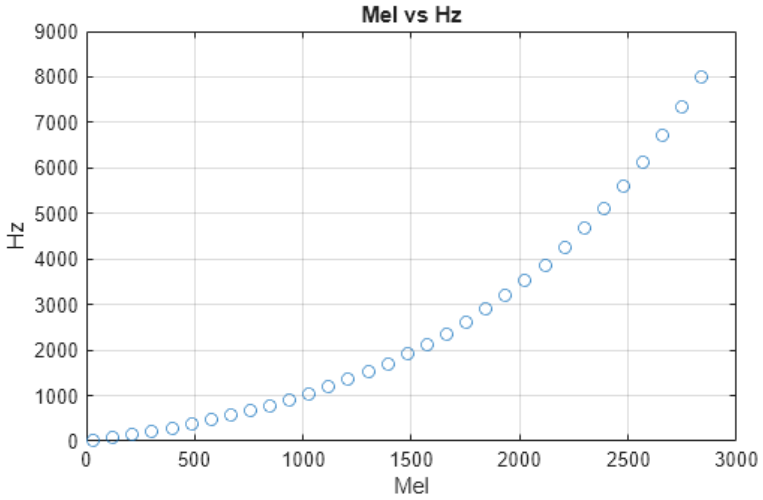
- Human hearing is less sensitive at higher frequencies
- The difference at low frequencies 50 Hz to 550 Hz is a massive change in perceived pitch
- However, the difference at high frequencies 13kHz to 15kHz is not a significant change in perceived pitch
- Information in low frequencies (like formants) is crucial for distinguishing vowels or nasals
- Modeling this human perceptual property improves speech recognition performance in the same way



Mel to the Rescue

- Designed to match human perception of pitch: how humans *perceive* sound at different frequencies
- Equal intervals in mels represent equal perceived distances between pitches to a human listener
- The scale is anchored at 1000 Hz being equal to 1000 mels
- Below approximately 500 Hz, the mel and Hz scales are roughly equivalent

$$mel(f) = 2595 \log(1 + \frac{f}{700})$$



Learn More

Slidev · Course Homepage