# COMP-4147

# Audio and Speech Processing: Introduction

**Instructor:** Dr. Poline XIAN

**Teaching assistants:**
Ms. SUN Yuxi
(23483628@life.hkbu.edu.hk)

# About me

- Lecturer: Dr. Poline XIAN
- Educational Background:
  - Ph.D., Duke University, USA, Signal Processing
- Work Experience:
  - Research scientist at Hong Kong Science and Technology Parks (HKSTP)
- Research Interests:
  - Audio processing, signal processing, high-dimensional data analysis
- Homepage: https://poline3939.github.io/
- Email: polinexian@comp.hkbu.edu.hk; or polinexian@hkbu.edu.hk
- Office: RRS 717

# In-class rules and appointment

- Turn off your cell phone or keep it on vibrate / silent mode
- Keep listening, thinking, and taking notes
- Interrupt me anytime if I am speaking too fast, or if you don't understand the concept
- Try to work on the in-class exercises
- Feel free to come to me during class breaks

- Email: send your questions to me or the TA.
- In-person meetings/discussions
  - After class, and
  - Monday: 2:00 pm-3:00 pm by appointment

# General course information

- Textbook: reading online
- Assessment:
  - Individual assignments (3 assignments, each 10%), 1 group project (30%), and final exam (40%)
  - Grade:

  0.3 x assignments + 0.3 x group project + 0.4 x exam

- To pass this course
  - Final exam score must be greater than 30%
  - Overall score must be greater than 35%
- Late submission will be penalized.

# Tentative schedule

| Lecture | Dates | Topics | Assignment |
|---|---|---|---|
| Lecture 1 | Jan 12-13 | Introduction to audio and speech processing | Assignment #1 Due: Feb 16 |
| Lecture 2 | Jan 19-20 | Basic concepts and techniques | |
| Lecture 3 | Jan 26-27 | Data mining and basic tools for audio processing (Lab) | |
| Lecture 4 | Feb 2-3 | Conventional and learning algorithms for acoustical modeling | |
| Lecture 5 | Feb 9-10 | Localization estimation and clustering with audio signals | Assignment #2 Due: March 16 |
| Lecture 5 (cont.) | Feb 16 | Lecture review and lab practice | |

# Tentative schedule

| Lecture | Dates | Topics | Assignment |
|---|---|---|---|
| Lecture 6 | Feb 23 – 24 | Audio application I: Sound separations and noise filtering | |
| Lecture 7 | March 2 - 3 | Audio application I (cont), Methods (DPRNN, TasNet) implementation (Lab) | Assignment #3<br>Due: April 13 |
| Lecture 8 | March 9 -10 | Audio application II: Audio recognition, music classification | |
| Lecture 9 | March 16 -17 | Audio application II (cont), methods implementation (Lab) | |

# Tentative schedule

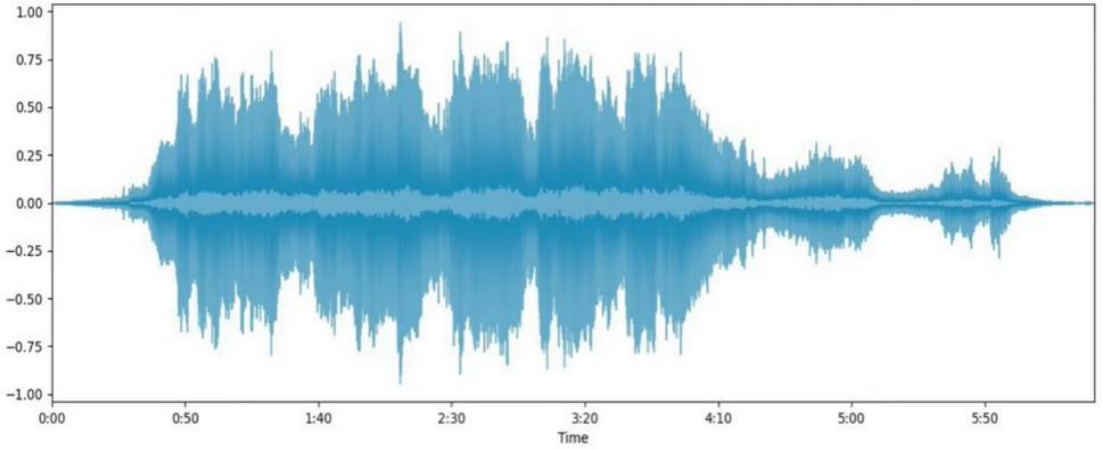
| Lecture | Dates | Topics | Assignment |
|---|---|---|---|
| Lecture 10 | March 23 – 24 | Audio application III: Audio generation and information retrieval | Group project<br>Due: April 20 |
| Lecture 11 | March 30 – 31 | Audio application III (cont), methods implementation (Lab) | |
| Lecture 12 | April 13 -14 | Advanced topics & Course Review | |
| Lecture 13 | April 21 | Presentation | |

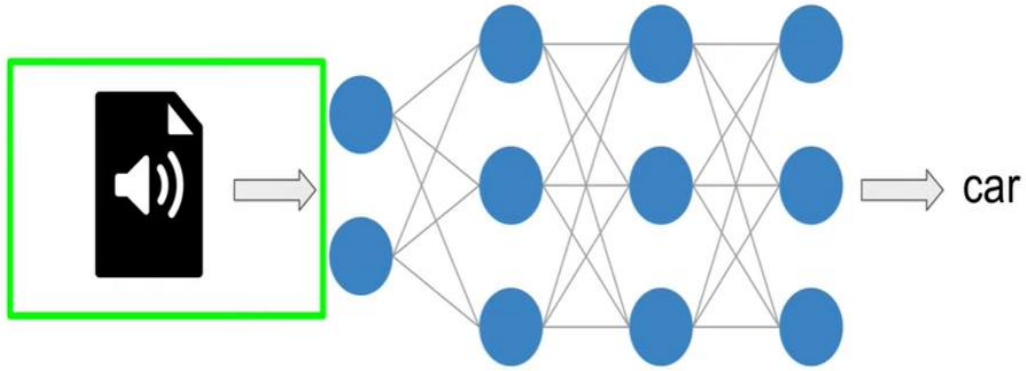
# What you will learn

- Get a deeper understanding of the audio signal
- Familiarize with frequency domain and time domain audio features
- Extract features from raw audio
- Recognize what audio features to use for Machine Learning applications
- Preprocess audio data for Machine Learning
- Understand (some) math behind audio transformations
- Use *librosa* for your audio projects.

# Applications

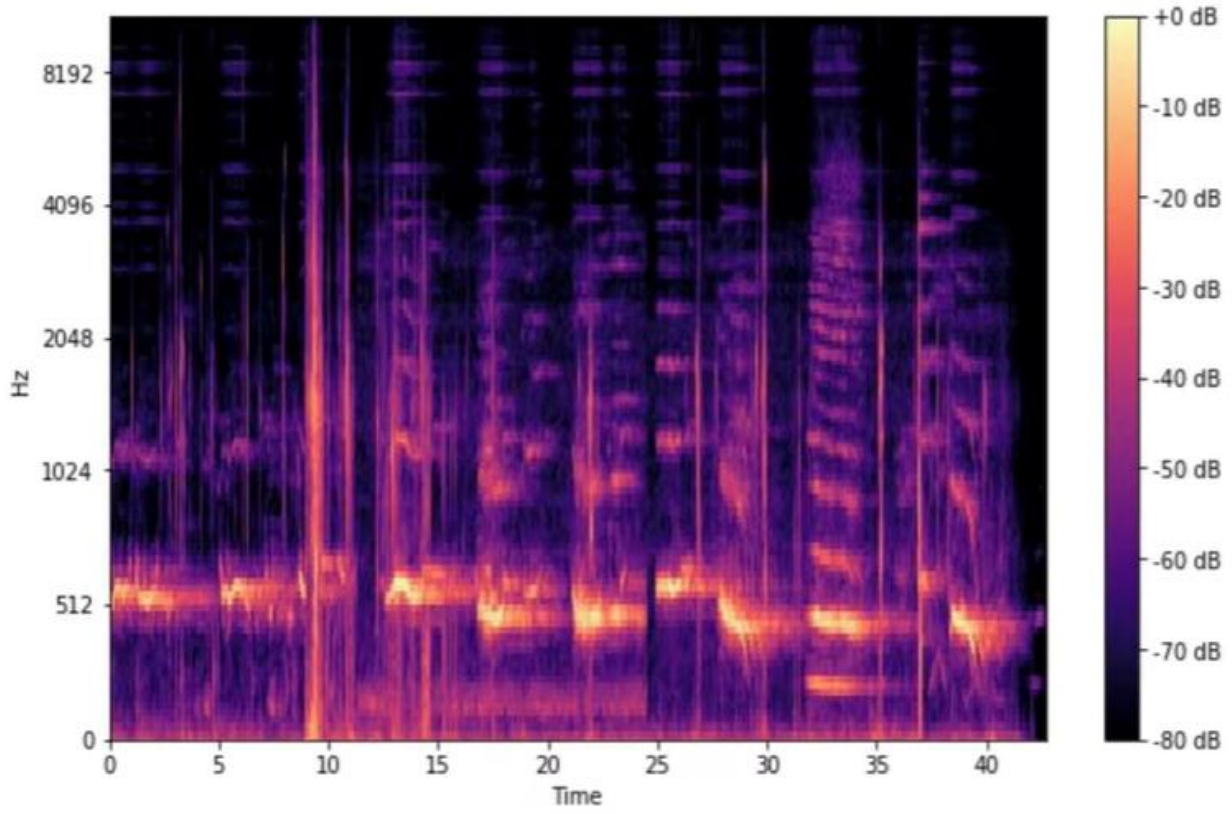
- Audio classification
- Speech recognition / speaker verification
- Audio denoising / audio upsampling
- Localization
- Music Information Retrieval
    - Music Instrument classification
    - Artist identification



# Content

- Sound waves
- DAC / ADC
- Time- domain and frequency-domain audio features (e.g, spectrogram)
- Audio transformation
  - Fourier Transform / STFT
  - Constant-Q Transform
  - Mel Spectrorams



# Recommended reading


- A. V. Oppenheim, and A. S. Willsky. Signals & Systems. Pearson Education, 2013. (Chapter 1-3, 5, 7)
- L. R. Rabiner, and R. W. Schafer, Introduction to Digital Speech Processing, Foundations and Trends in Signal Processing 1 (1-2), 1-194, 2007.
- K. P. Murphy, Probabilistic Machine Learning: Advanced Topics, MIT Press, 2023.
- D. Jurafsky, and J. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd ed. 2024. (Chapter 9, 25, 26)
- I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016.
- K. Gröchenig, Foundations of time-frequency analysis. Springer Science & Business Media, 2001. (Chapter 3, 5, 10)
- F. Chollet, and F. Chollet. Deep learning with Python. Simon and Schuster, 2021.
- Selective reading from recent issues of IEEE Transactions on Pattern Analysis and Machine Intelligence, Transactions on Signal Processing, Proceedings of ICASSP, Interspeech, and ICML.

# Beware

- Unless otherwise stated, all work submitted by you should be your own.
- Copying or sharing of assignments or any submitted work is cheating.

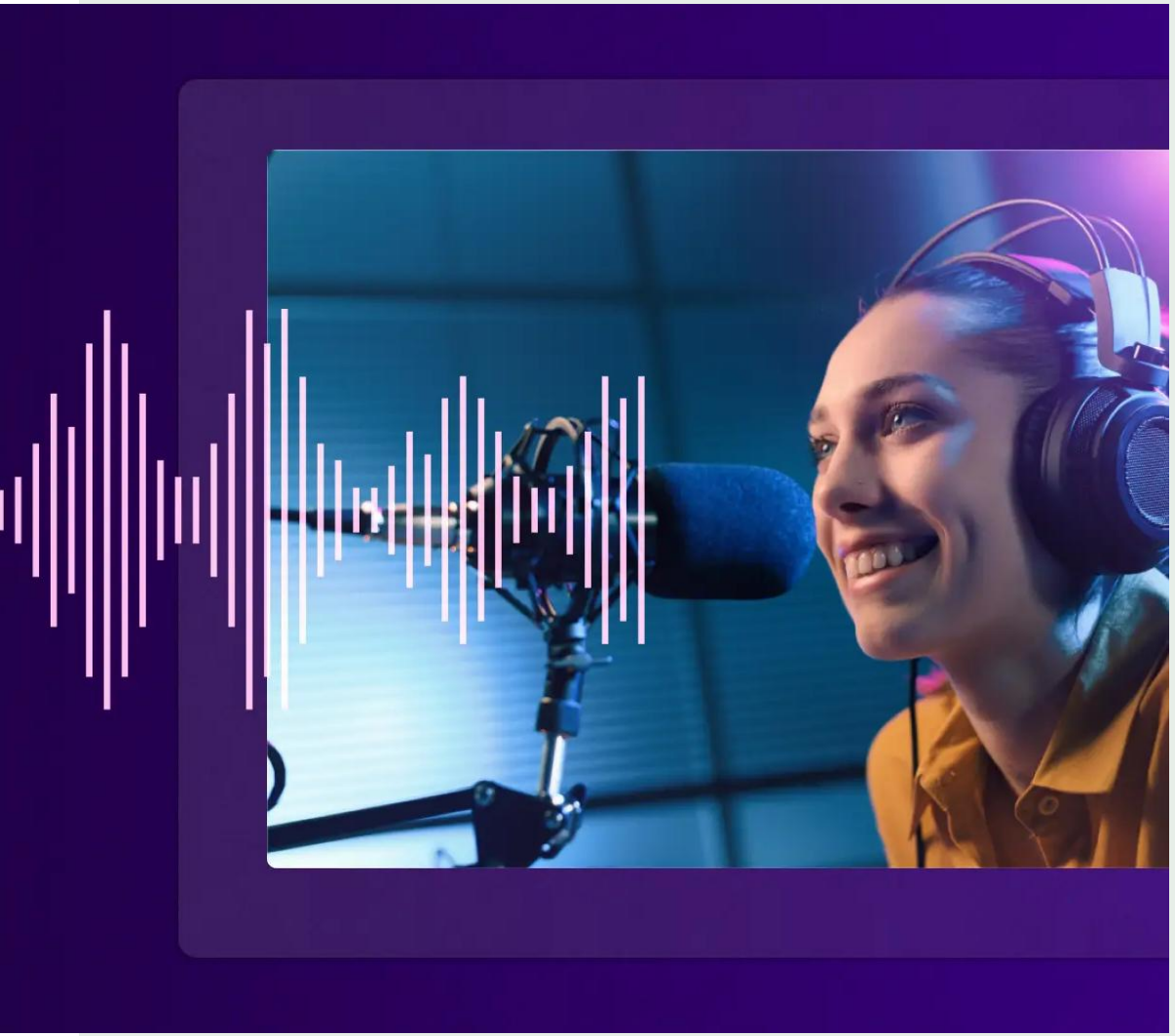
## Cost of Plagiarism

- Plagiarism is an offense and will result in appropriate disciplinary action against those involved.
- Penalty will be applied indiscriminately among those who are involved (the one who copies and the one being copied). The minimum penalty would be receiving zero marks for the submitted work.
- Please refer to the following URL for the university's guidelines on penalizing plagiarism: https://bba.hkbu.edu.hk/academics/teaching-and-learning- supports

# Use of Generative AI

- Proper Use of Generative AI
  - Explaining or clarifying concepts;
  - Demonstrating and guiding practices of techniques;
  - Planning and brainstorming on projects;
  - Giving feedback on drafts;
  - Generating samples for discussion and critical review.

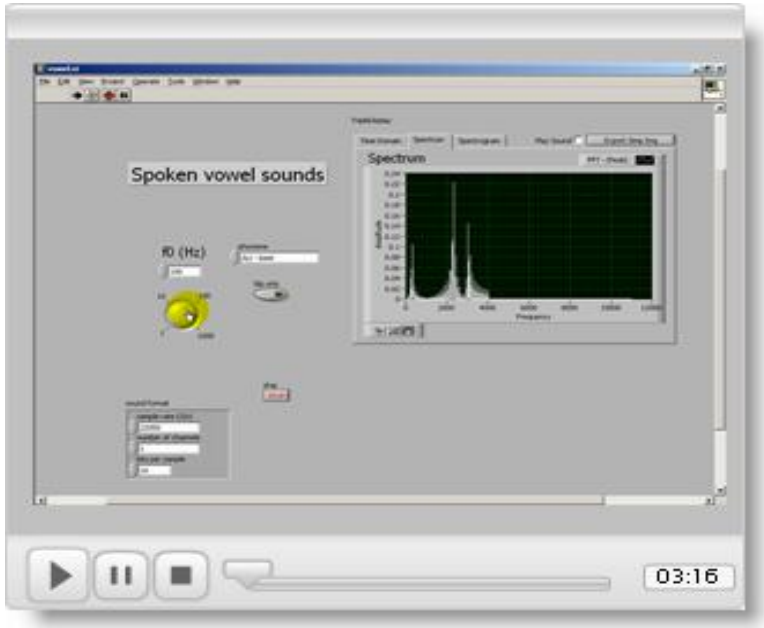
- University's guideline: https://bba.hkbu.edu.hk/academics/teaching-and-learning- supports



# Overview

# History of Audio and Speech Processing

- Early Foundation (1930s – 1950s):
  - Analog Synthesis: Homer Dudley at Bell Labs introduced the Voder (Voice Operating Demonstrator) at the 1939 for synthesizing speech and the Vocoder for voice compression




Information Theory: Claude Shannon and Harry Nyquist developed theories for sampling and digital communication (PCM), crucial for digital audio.

In **1952**, *Bell Lab* designed the "**Audrey**" system, which could recognize a single voice speaking **digits** aloud

# History of Audio and Speech Processing

- Digitalization & Statistical Modeling (1960s – 1990s):
  - LPC and Coding: Linear Predictive Coding (LPC) became key for speech, while DCT (1974) and MDCT (1987) led to MP3.
  - Computer Music: Max Mathews synthesized the first computer audio in 1957.
  - Speech Recognition Research: Projects like Sphinx (CMU) introduced continuous speech recognition
  - Text-to-Speech (TTS): The Speak & Spell (1978) used phoneme-based synthesis, and early PC software like SAM (1979) brought speech to home computers.

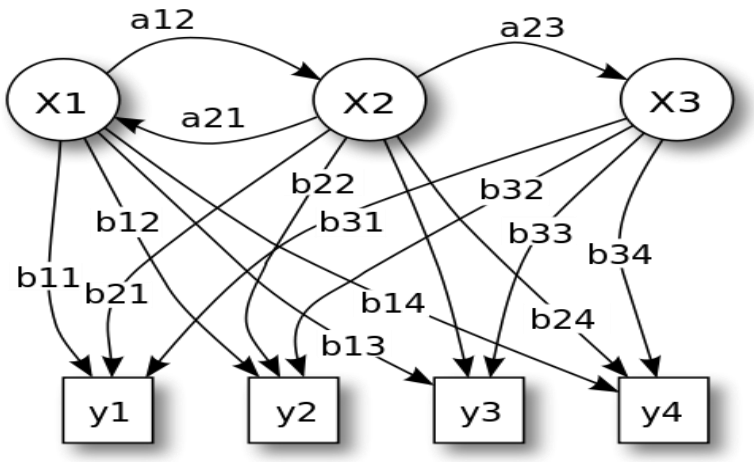
In **1962,** *IBM* introduced "**Shoebox**" which understood and responded to **16 words** in English.

https://web.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/354_LALI-ASRHistory-final-10-8.pdf

# History of Audio and Speech Processing

The system of the Defense Advanced Research Projects Agency (DARPA) was capable of understanding over **1,000** words. **Siri** was a spin-out of DARPA development.

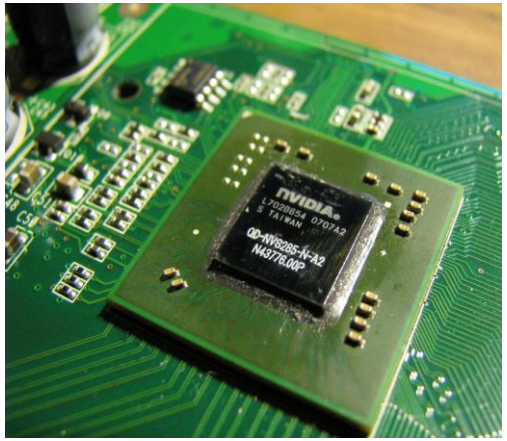
The ‘80s saw speech recognition vocabulary go from a few hundred words to **several thousand words,** thanks to **HMM**



# History of Audio and Speech Processing

- Modern Era (2000s - Present)
  - Neural networks: deep learning significantly improved accuracy, powering modern systems.
  - AI assistants: Voice search and assistants become common. Research in realistic TTS (e.g., Google WaveNet).
  - Advanced synthesis: Adobe's Voco (2016) allows text-to-speech from minimal training data, enabling voice editing and restoration.

**Guess which one is generated?**





Speech recognition has been propelled forward since the 00s in large part because of **faster processors.** Next is the era of big data, machine learning, and GPUs

# Task: Audio Separation


Mixture separation and enhancement

- Female mixture original
  - Speaker 1
  - Speaker 2
- Male mixture original
  - Speaker 1
  - Speaker 2

Music Separation

Mixture Piano Drums Guitar

# Task: Audio Enhancement (Google Research)



Test demo 1:

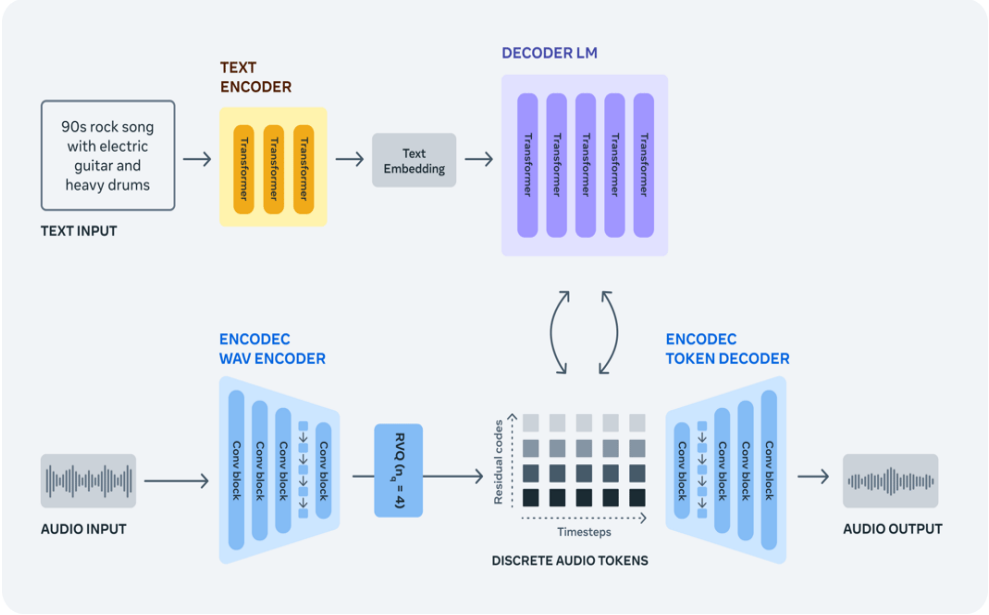
Test demo 2:

Improvement:

Improvement:

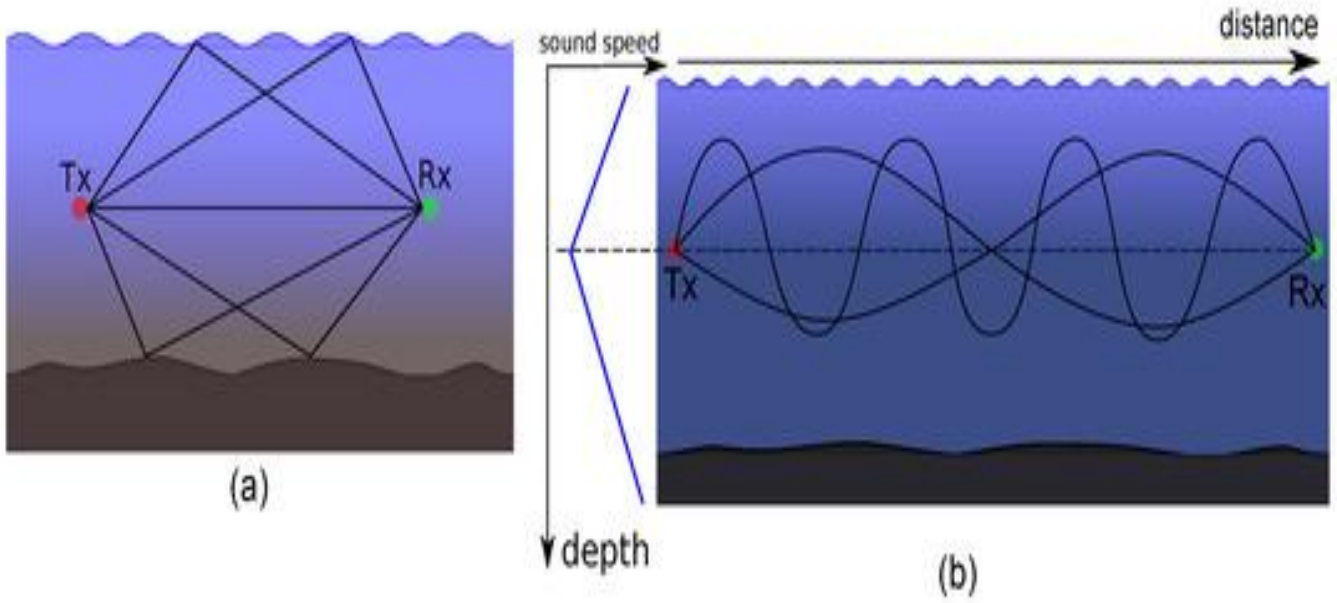
# Task: Audio Generation

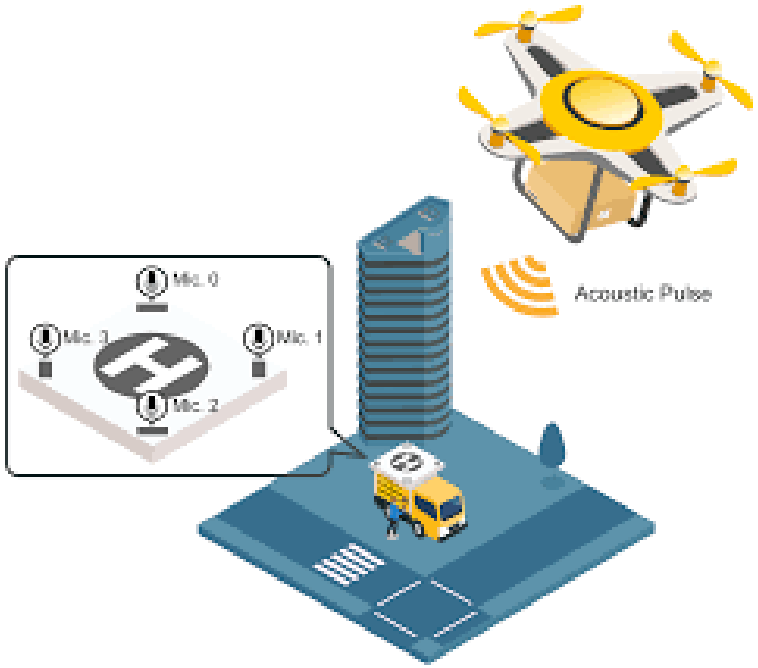

Facebook Research demo of voice plus different environments (AI AR/VR)




- [MusicGen: https://audiocraft.metademolab.com/musicgen.html](https://audiocraft.metademolab.com/musicgen.html)

# Tasks: Sound Localization



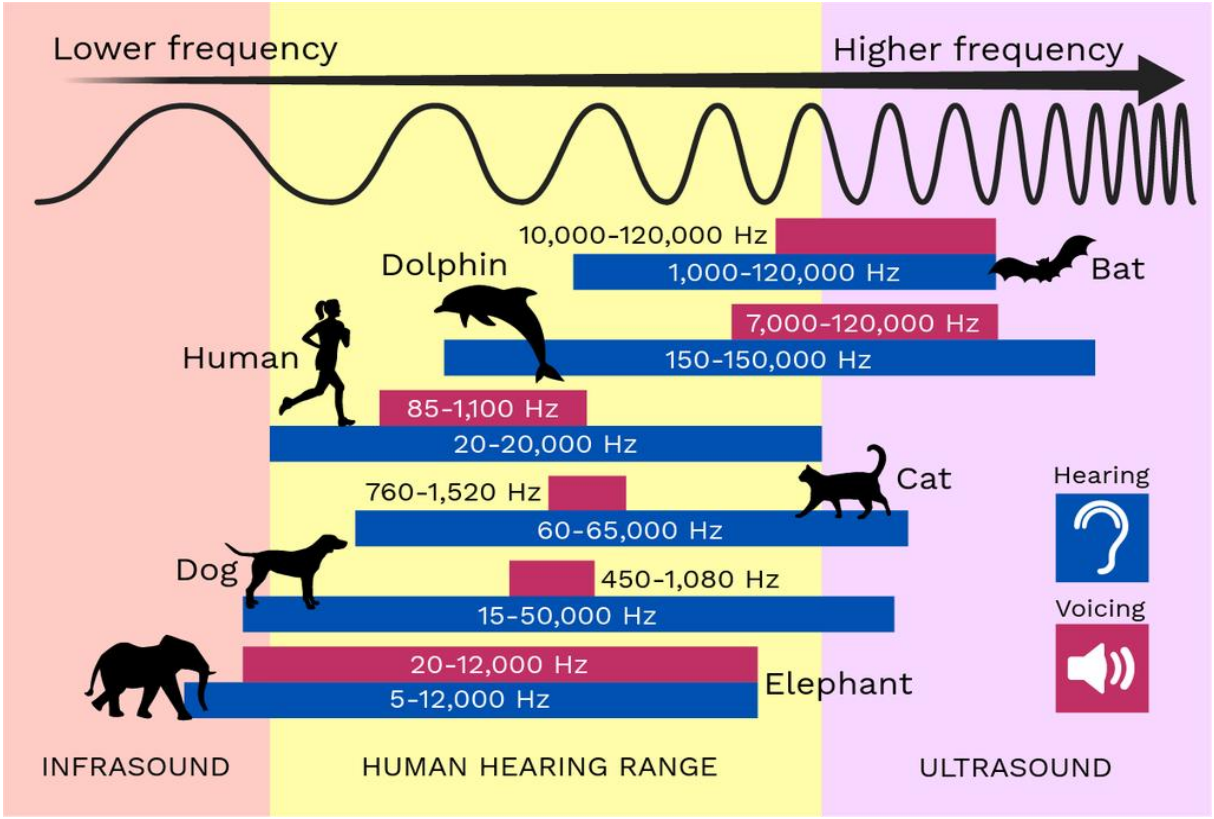




- When there are obstacles that block light and obscure vision, the audio system provides an advantage.
- Vision is ineffective in darkness and underwater, but sound localization ability remains.

# Sound and Waveform

# Audio signal

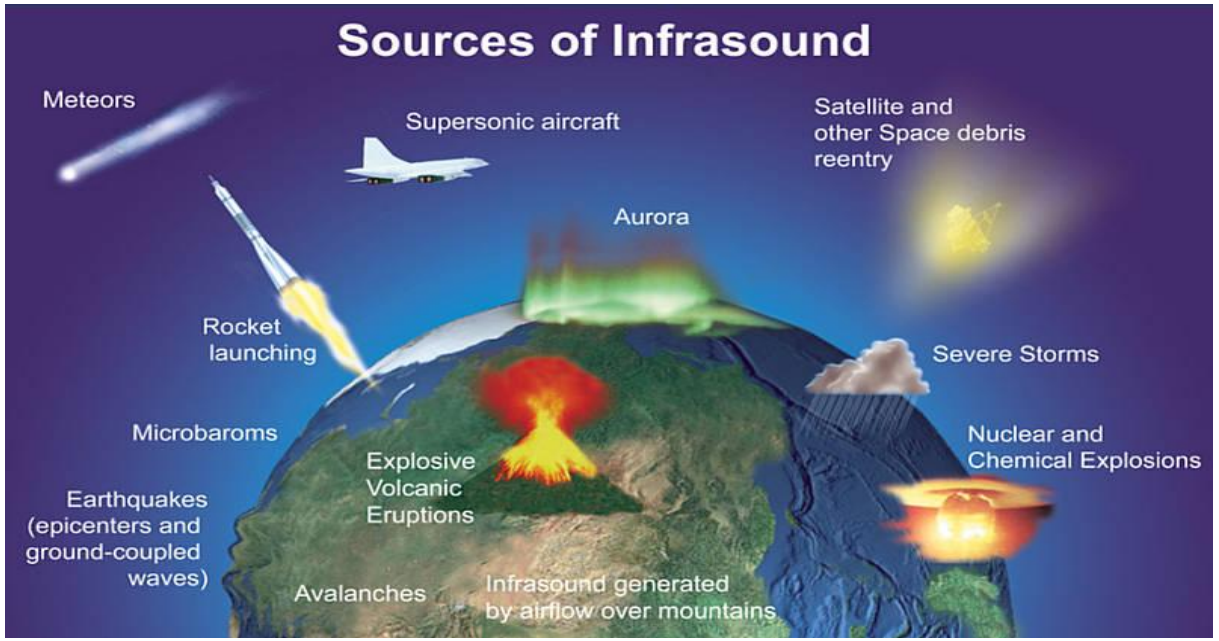
- Audio signals are the representation of sound, which is in the form of digital and analog signals
- Sound frequencies range between 20 Hz and 20,000 Hz. It is the lower and upper limit of our ears.
- The human auditory system is most sensitive between 500 Hz and 5000 Hz.




Hearing and Voicing Range

# Audio signal

- Sounds **below the frequency of 20 Hz** can also affect us even though we cannot hear them.
- **Infrasound** at the frequency of 7 Hz is dangerous. The frequency is close to the characteristic frequencies of our organs, such as the heart and brain.
- **Characteristics**: long wavelengths, can travel long distances, penetrates obstacles well.
- **Examples**: communication in elephants, monitoring earthquakes/volcanoes, and wind turbines that cause discomfort to humans.
- **Application**: Long-range sensing.

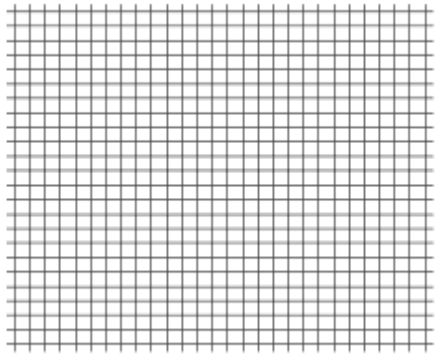



# Ultrasound

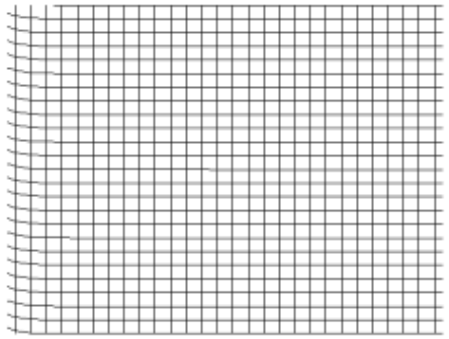
- Frequency: above 20,000 Hz (20 kHz)
- Characteristics: short wavelengths, reflection of surfaces, good for detailed imaging.
- Uses: Bats and dolphins use it for echolocation, medical scans (sonograms), breaking kidney stones, sonar, and non-destructive testing (finding cracks).
- Application: Detailed imaging and short-range navigation or detection.

# Sound Production

- Sound is a vibration that propagates as an **acoustic wave** through a transmission medium such as a gas, liquid, or solid.
- It propagates as *longitudinal waves* and as *transverse waves*.
- Acoustics is the interdisciplinary science that deals with the study of mechanical waves in gases, liquids, and solids, including vibration, sound, ultrasound, and infrasound.
- Sound waves are often described in terms of sinusoidal plane waves. They are determined by: frequency or wavelength, amplitude, sound pressure or intensity, sound speed, and direction.



Longitudinal plane wave

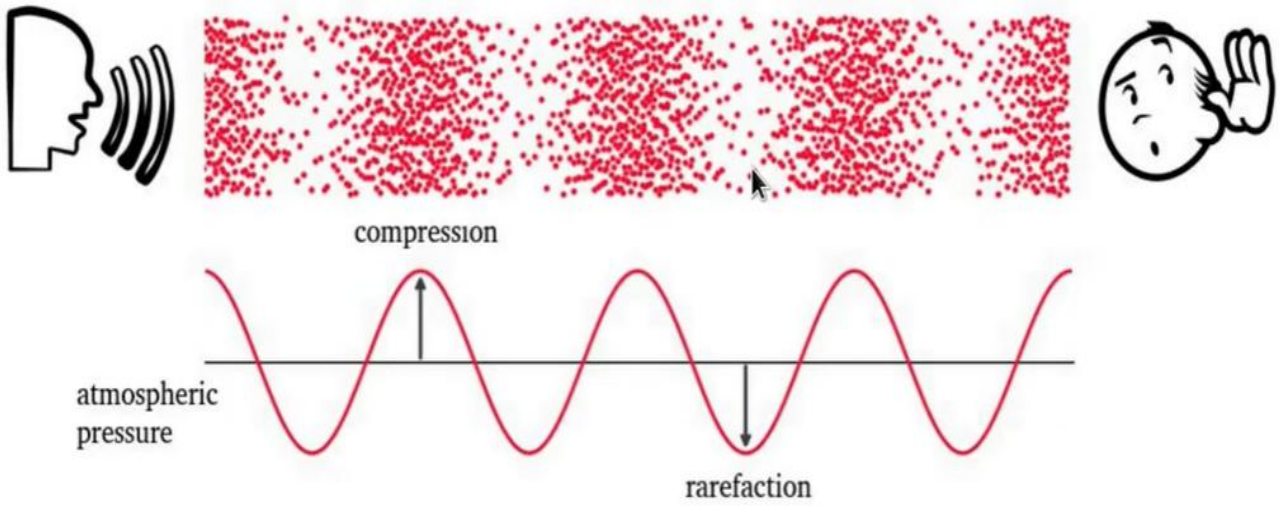


Transverse plane wave

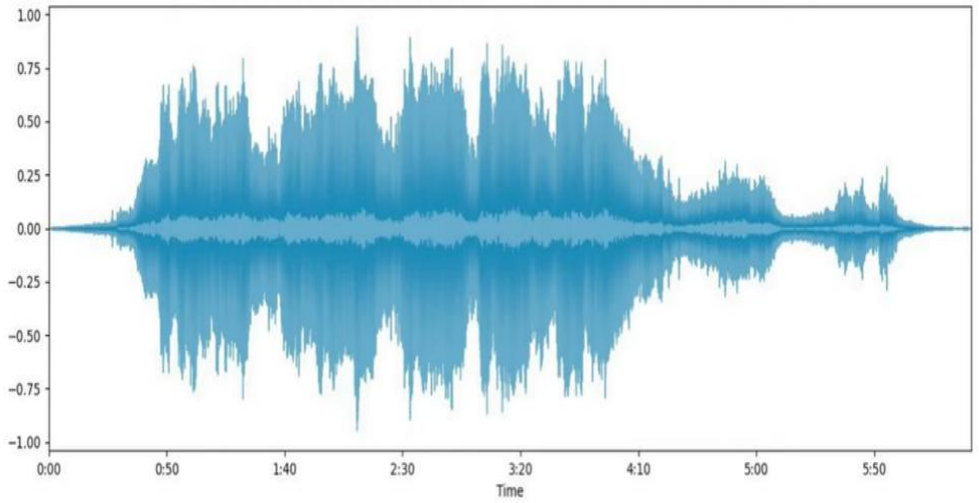
Acoustic wave equation:

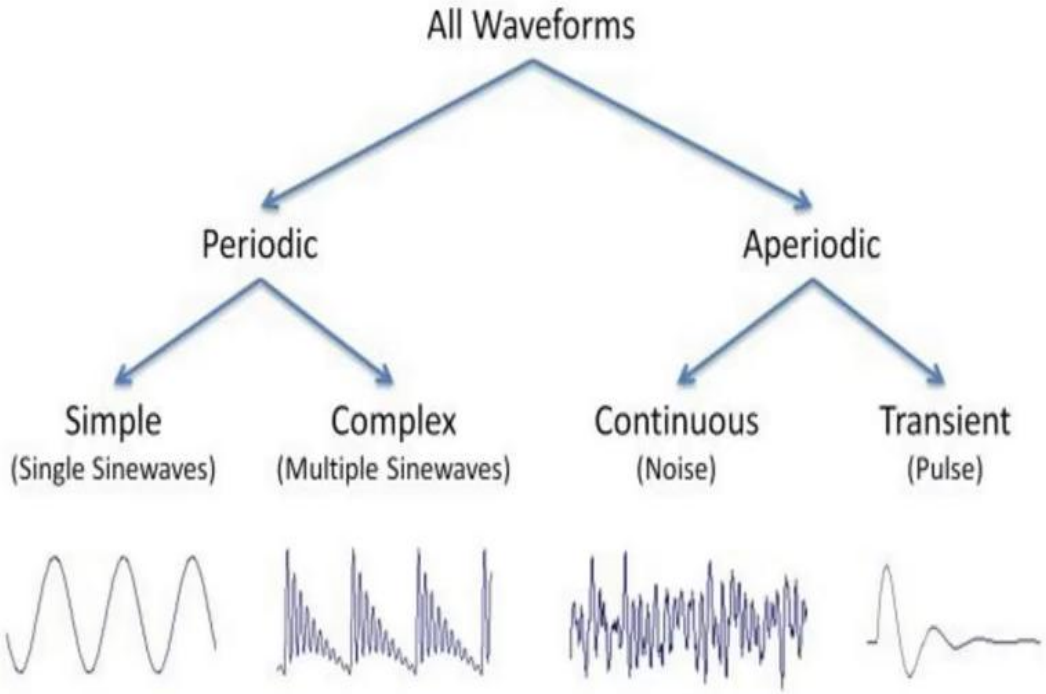
$$p_{xx} - \frac{1}{c^2} p_{tt} = 0,$$

# Visualization of sound wave



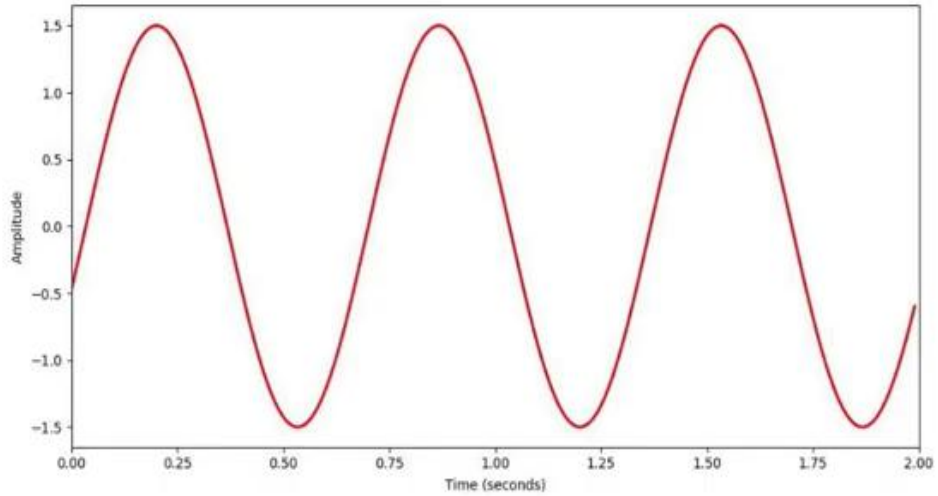

# Waveform

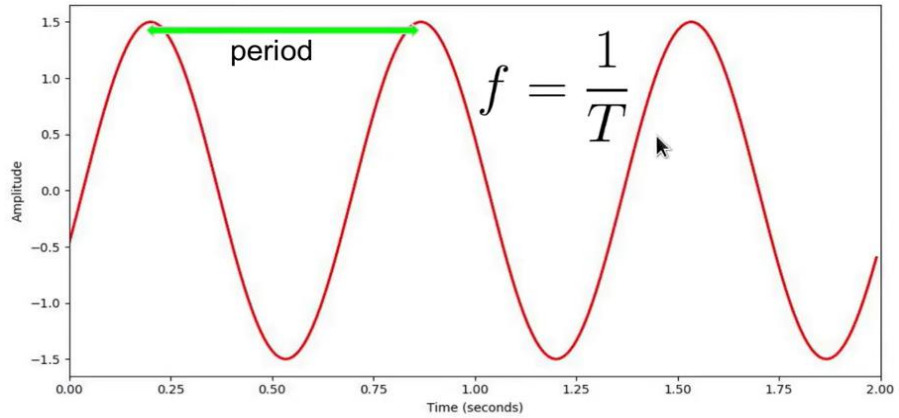



Periodic and aperiodic sound

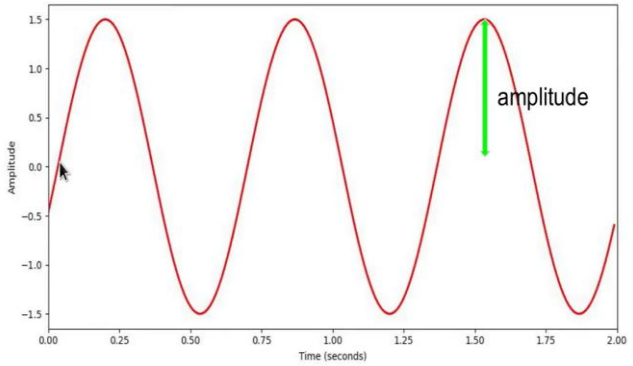
# Waveform



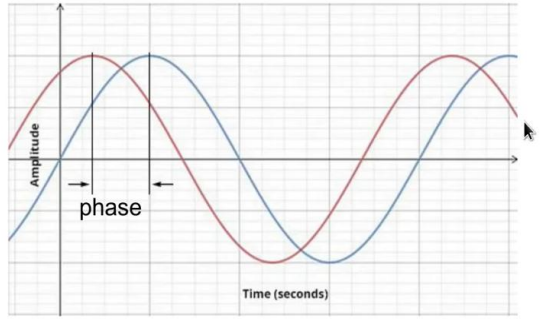




- Period (T): the time for one cycle
- Frequency (f): how many cycles occur in a given time

$$y(t) = A\sin(2\pi ft + \varphi)$$

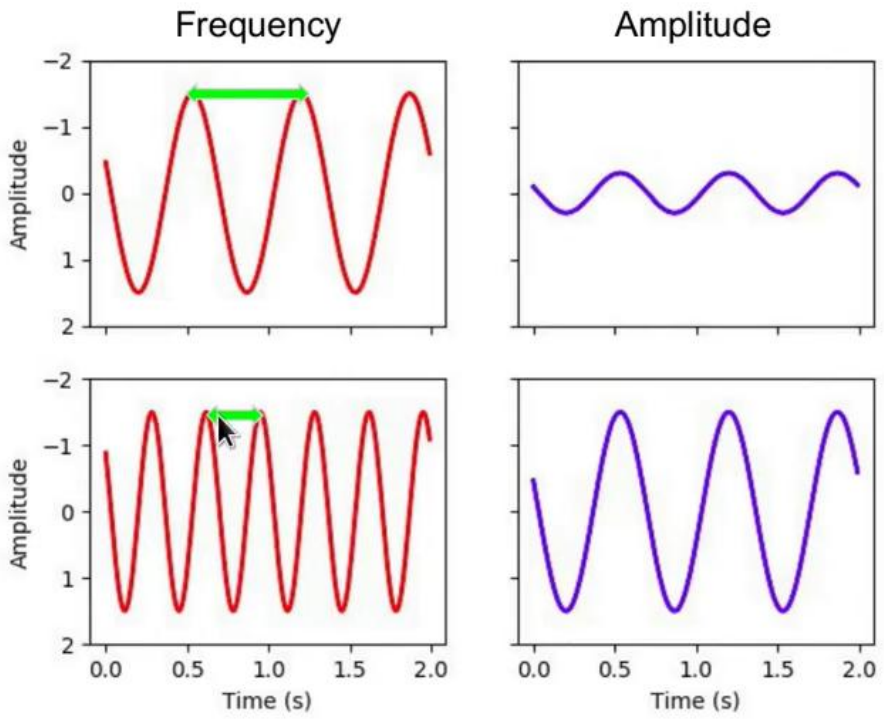



Amplitude (A): The wave's height from its center

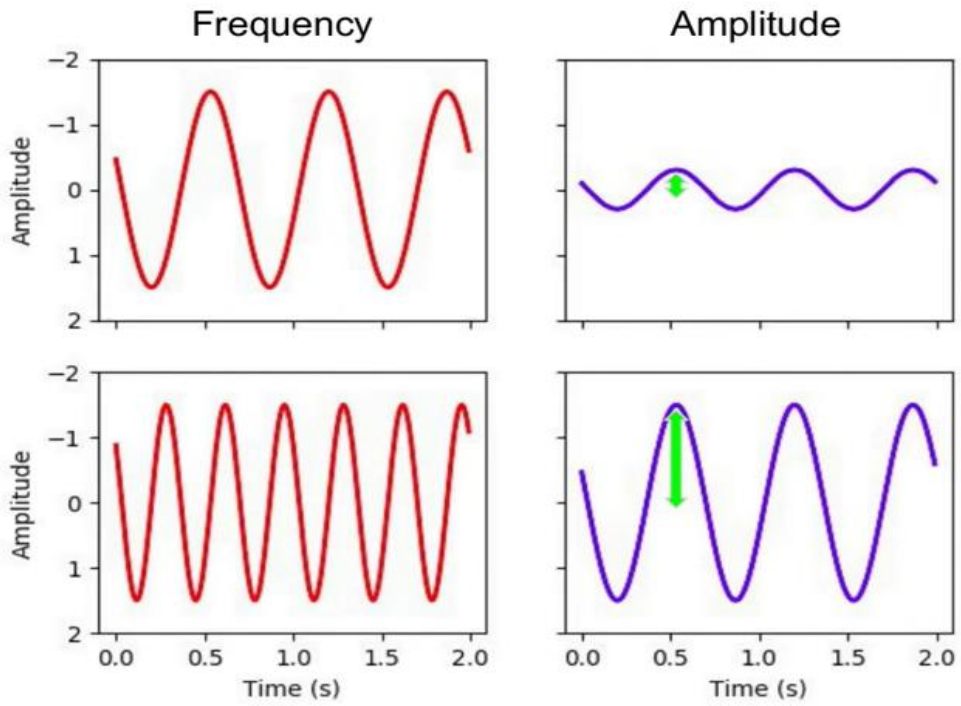



Phase ($\varphi$): The horizontal shift from the standard position.

# Frequency and amplitude



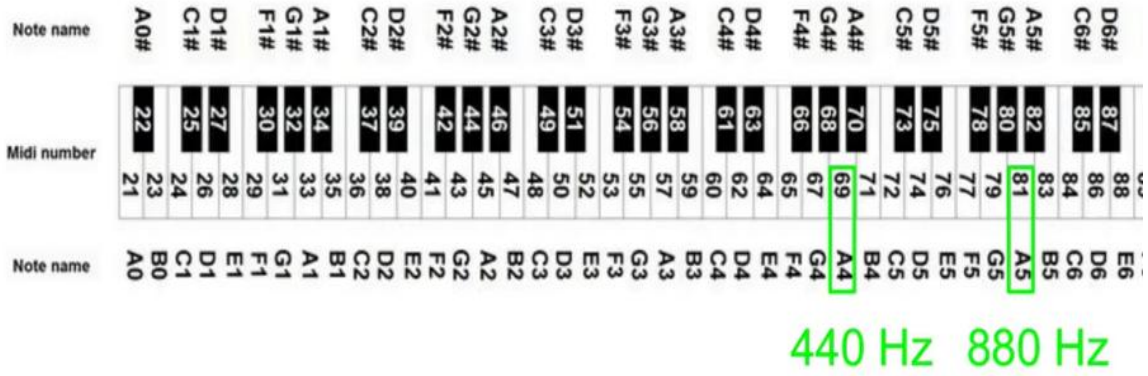

Higher frequency -> higher sound

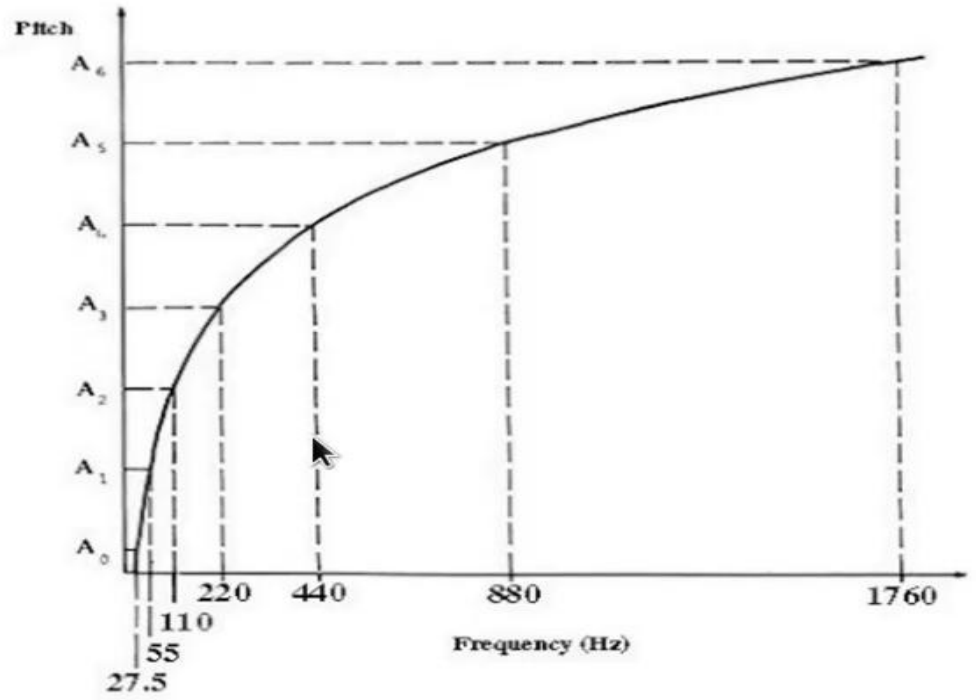



Larger amplitude -> louder

# Pitch

- The way we hear or perceive frequency
- Logarithmic perception
- 2 frequencies are perceived similarly if they differ by a power of 2.



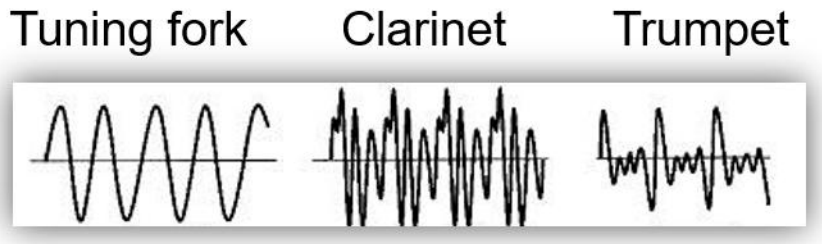

Mapping pitch to frequency:

$$F(p) = 2^{\frac{p-69}{12}} \cdot 440$$

$$F(60) = 2^{\frac{60-69}{12}} \cdot 440 = 261.6$$

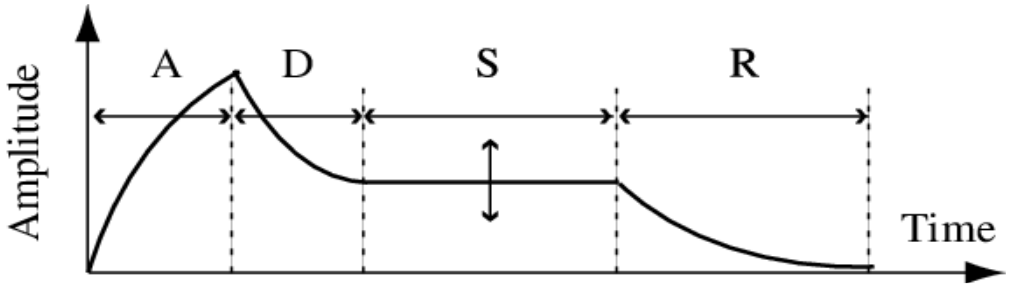
# Timbre

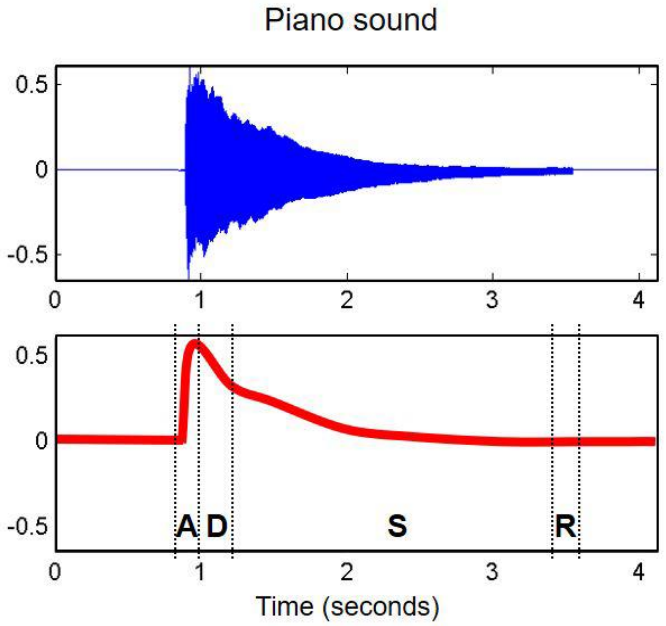
- Color of sound
- Difference between two sounds with the same intensity, frequency, and duration
- Described with words like bright, dark, dull, harsh, warm
- Features
  - Multidimensional
  - Sound envelope
  - Harmonic content
  - Amplitude and frequency modulation

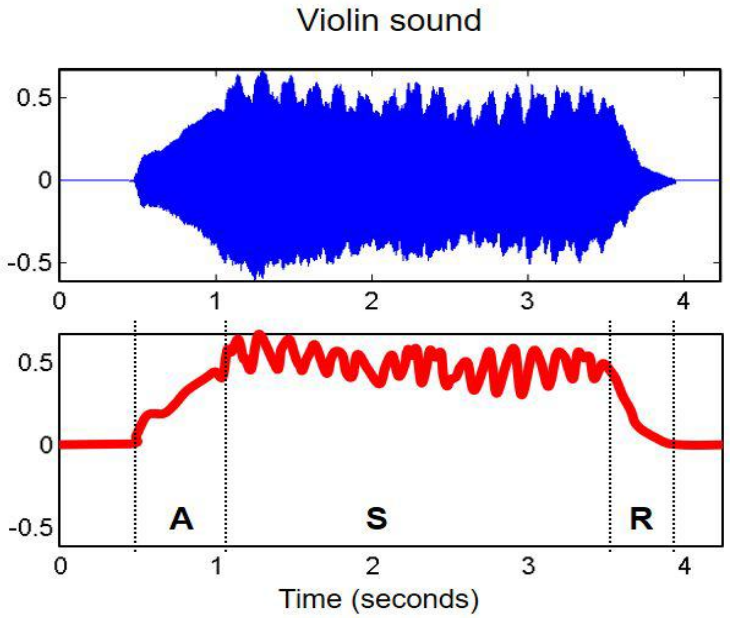


# Sound envelop

- It describes how the sound amplitude (volume) changes over time.
- Attack-Decay-Sustain-Release Model




- Attack (A): the time it takes for the sound to rise from silence to its maximum volume
- Decay (D): The time it takes for the sound to drop from the peak to the sustain level
- Sustain (S): The relatively constant volume level the sound maintains
- Release (R): The time for the sound to fade from sustain level to silence.

# Complex sound

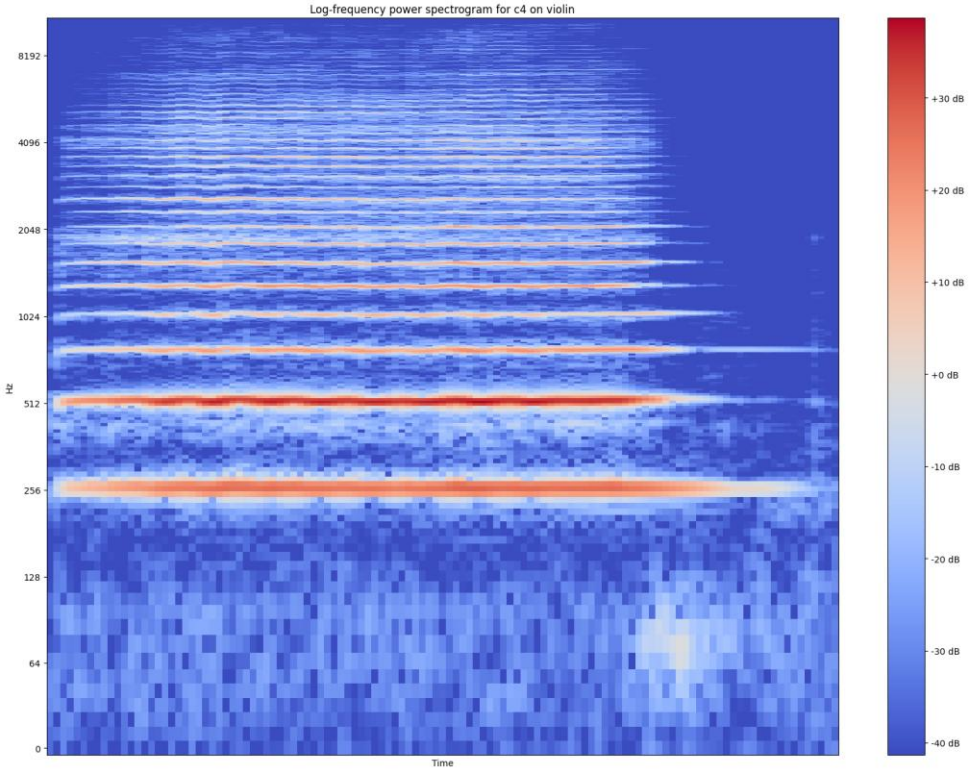
- Superposition of sinusoids
- A *partial* is a sinusoid used to describe a sound
- The lowest partial is called the *fundamental frequency*
- A harmonic partial or overtone is a frequency that's a multiple of the fundamental frequency

$$f_1 = 440 \text{ Hz}, f_2 = 2 \times 440 = 880 \text{ Hz},$$

$$f_3 = 3 \times 440 = 1320 \text{ Hz}, \ldots$$

- Inharmonics: overtones that are not integer multiples, often creating noisy sounds.

Violin



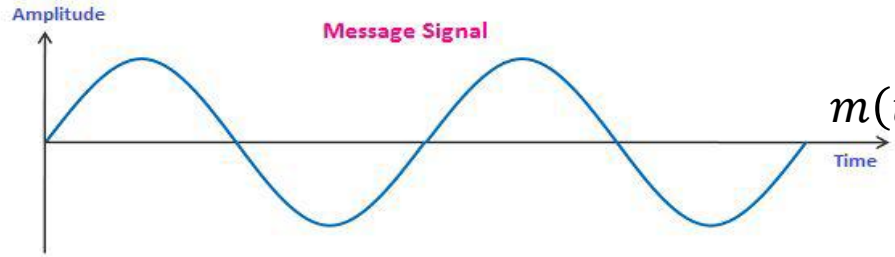

Distribution of energy across partials

https://colab.research.google.com/drive/1PYxitw5MB1wOFifDI_jkWgsBIFiy4cpJ

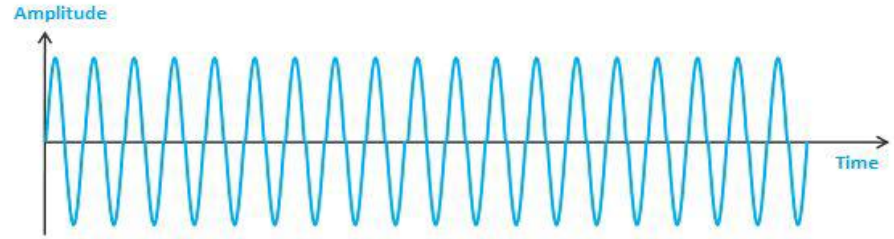
# Harmonic vs inharmonic instruments

# Frequency modulation

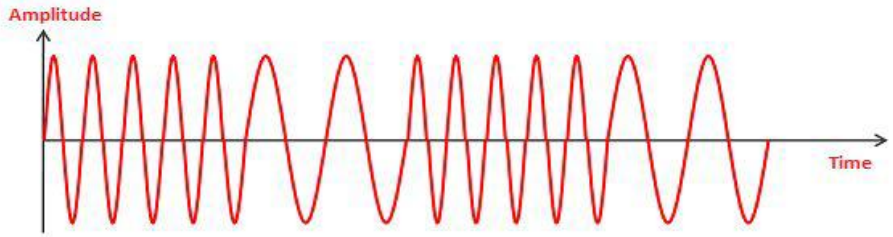
- Periodic variation in frequency
- AKA *vibrato* in music




$$m(t) = A_m \cos(\omega_m t)$$




$$A_c \cos(\omega_c t)$$

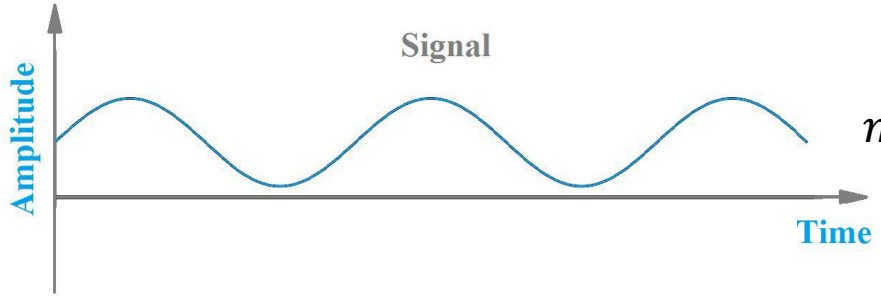



$$s_{FM}(t) = A_c \cos(\omega_c t + \emptyset(t)),\ \emptyset(t) = k_f \int m(t)dt$$

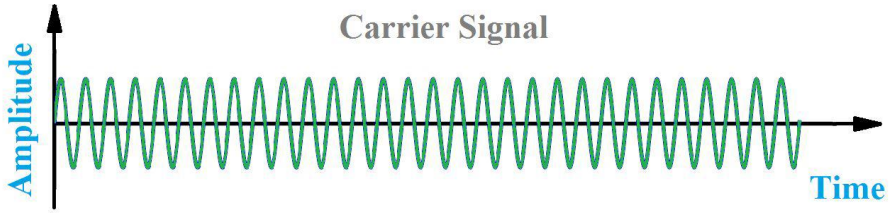
https://www.youtube.com/shorts/mCVGexw_Iqc

# Amplitude modulation

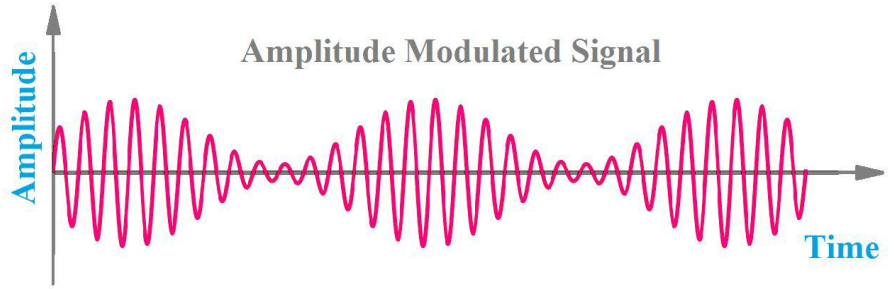
- Periodic variation in amplitude
- AKA *tremolo* in music




$$m(t) = A_m \cos(\omega_m t)$$




$$A_c \cos(\omega_c t)$$




$$s_{AM}(t) = [A_m \cos(\omega_m t) + A_c] \cos(\omega_c t)$$

# Sound recap

- Sound is a wave
- Frequency, period, amplitude, and phase
- Pitch, and timbre

# Review questions

- Describe the ADSR (Attack-Decay-Sustain-Release) model for a sound envelope. How does timbre differ between a violin (harmonic) and a drum (inharmonic)?

- The human hearing range is 20 Hz to 20,000 Hz. Identify and describe two applications: one for infrasound (20,000 Hz).

- A sound wave has a period $T = 0.2$ seconds and amplitude $A = 0.5$. Calculate the frequency $f$. Sketch a simple waveform and label period, frequency, amplitude, and phase.

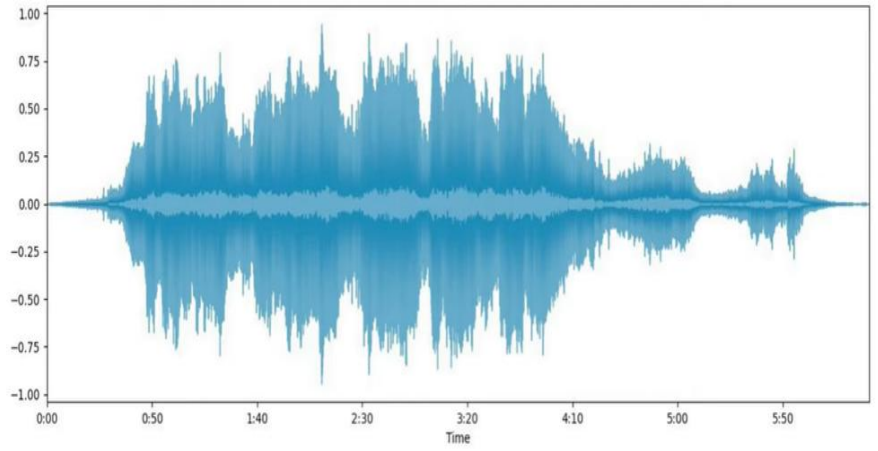
- Using the formula $F(p) = 2^{\frac{p-69}{12}} \cdot 440$ ,calculate the frequency for pitch $p = 60$(middle C). If a note is one octave higher ($p = 72$), what is its frequency?

# Time Domain and Frequency Domain

# Time domain vs Frequency domain

Time domain

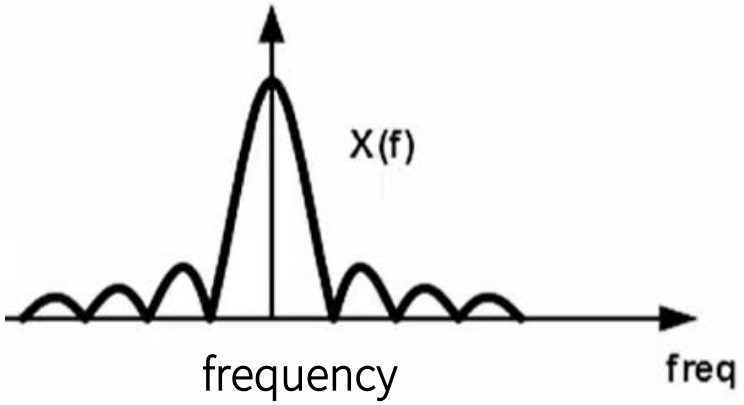
- The real world happens in the time domain.
- The independent variable is time, denoted by *t*




time

Frequency domain

- Signals may be represented by frequency components.
- Independent variable is frequency *f*. You may use the angular frequency $\boldsymbol{\omega = 2\pi f}$



# Time domain vs Frequency domain

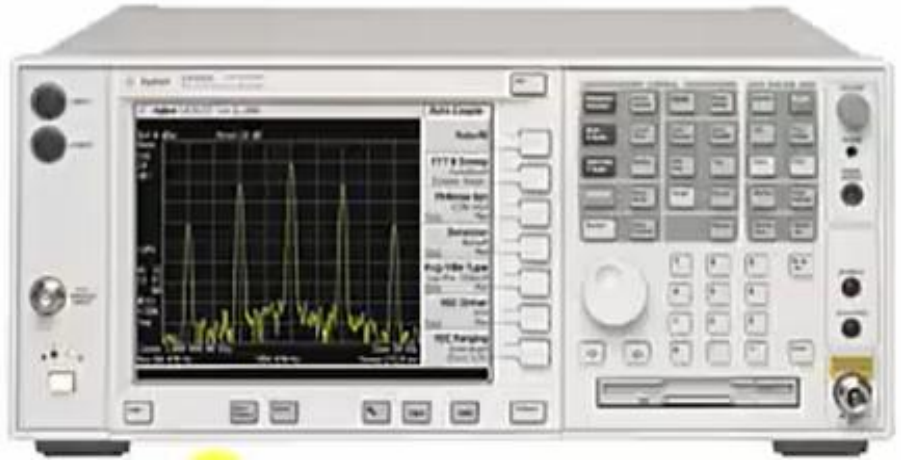
Measuring signal in the time domain.

- Oscilloscope: time

time

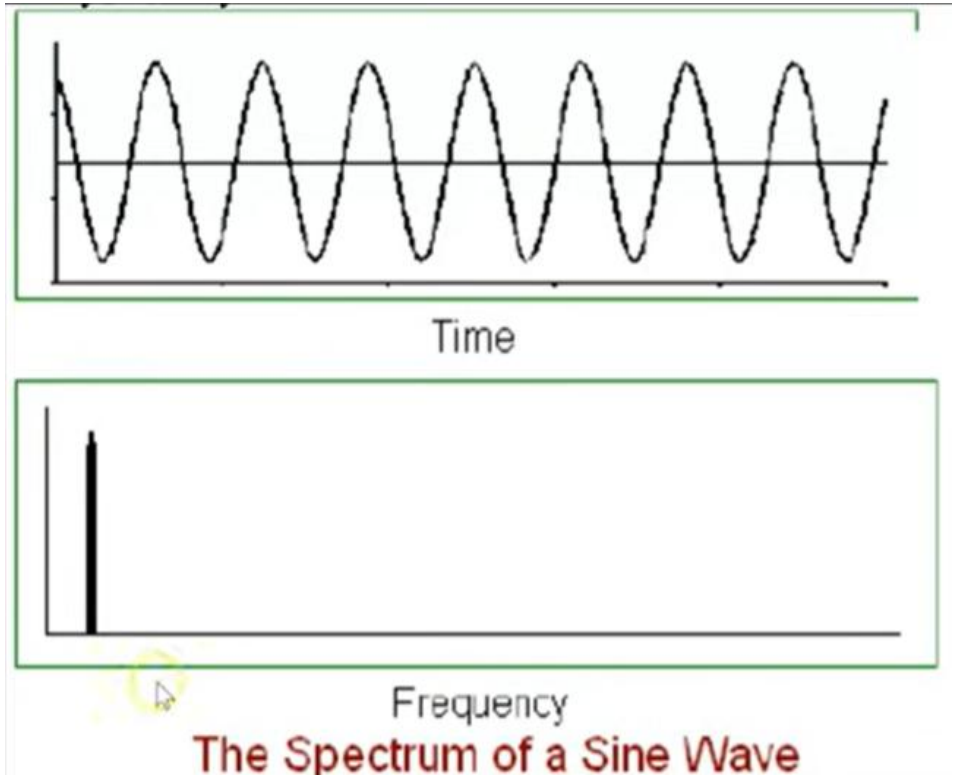
Measuring signal in the frequency domain.

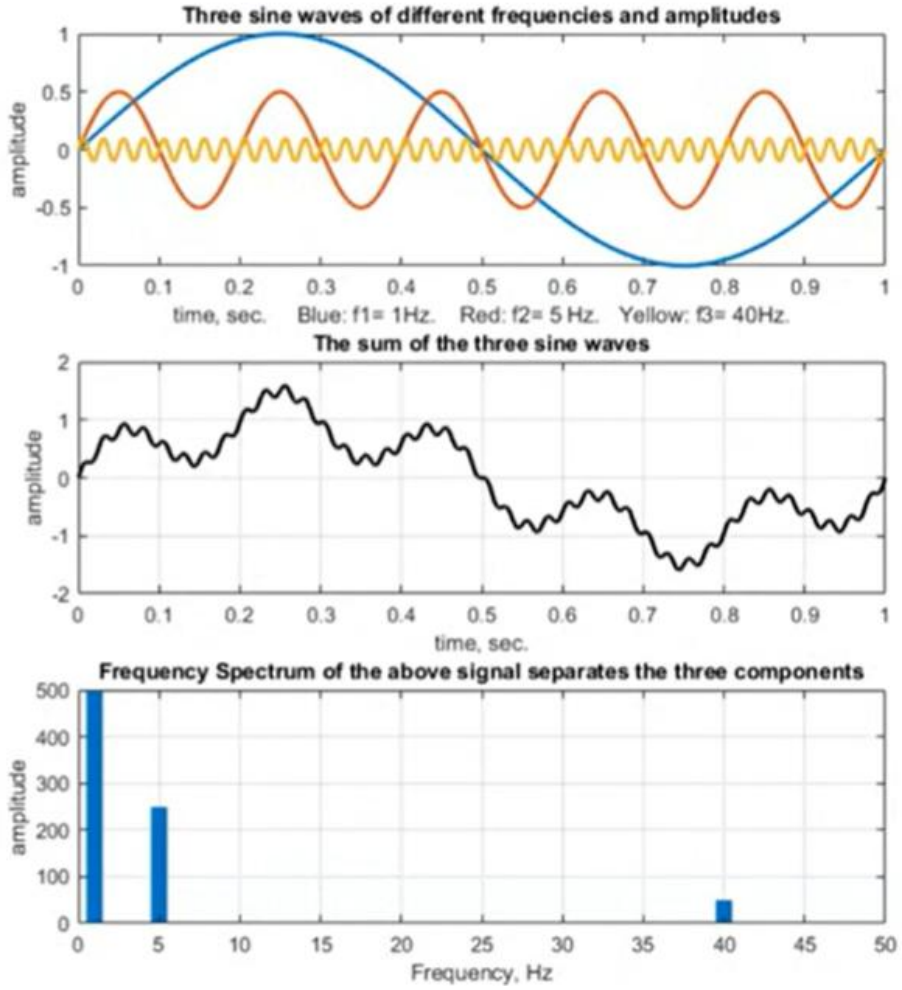
- Spectrum analyzer: frequency



frequency

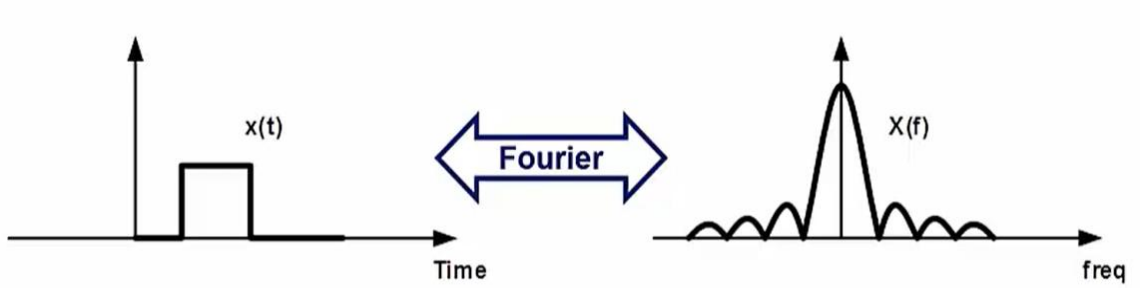
# Time domain vs Frequency domain



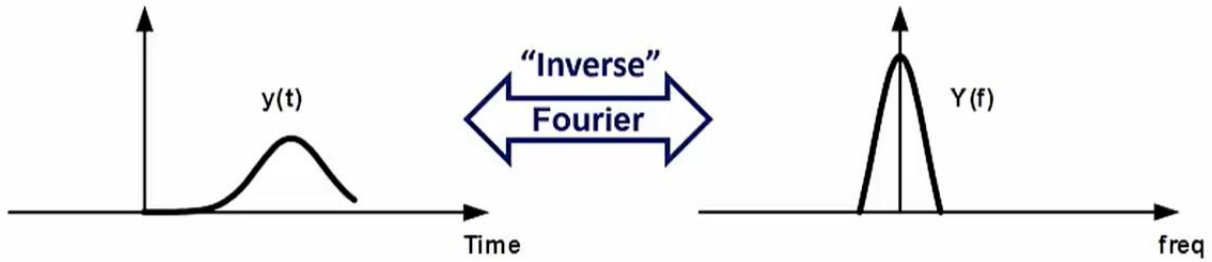



# Time domain vs Frequency domain

- The two representations are related through the Fourier transform, and both contain the same information.
- By changing the signal in one domain, its representation in the other domain changes as well.



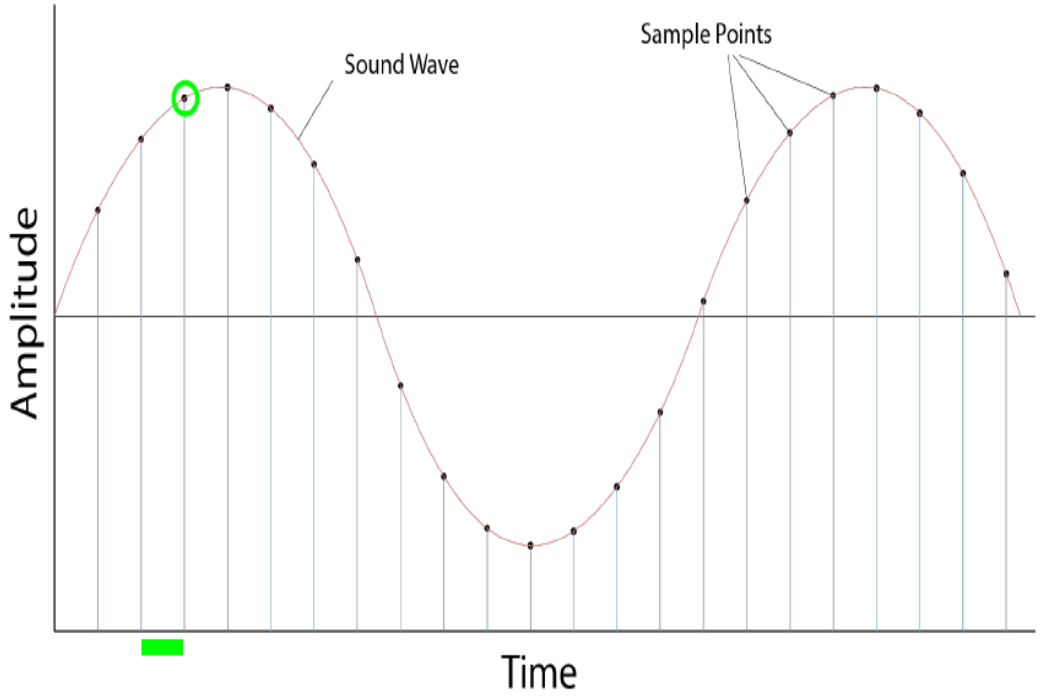

$$x(t) \quad \text{Fourier} \quad X(\omega)$$




$$\hat{x}(f) = \int x(t) e^{-i2\pi f t}\, dt \quad \Rightarrow \quad \hat{x}(f) = \sum_n x(n)\, e^{-i2\pi f n}$$

discretized

# Digitization




- $t = nT$
- $x(t)$ become $x(n), n = 1, \dots, N$, finite samples

$$\hat{x}(f) = \int x(t) e^{-i2\pi f t}\, dt$$

$$\hat{x}(f) = \sum_n x(n)\, e^{-i2\pi f n}$$

Considering frequency is finite, *N* samples in total

- $\hat{x}(f) = \sum_n x(n)\, e^{-i2\pi f n}$ becomes

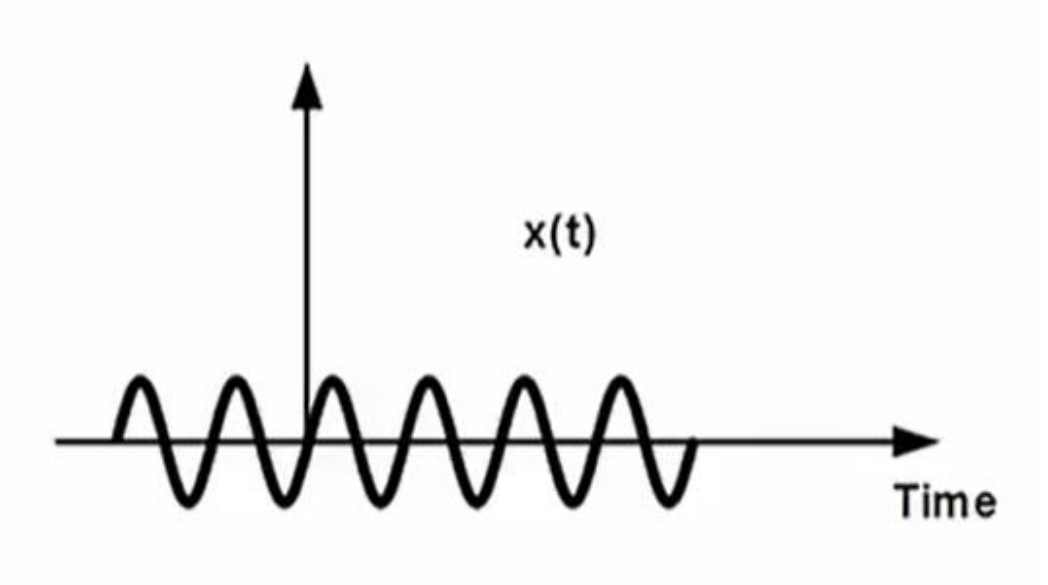
Discrete Fourier Transform (DFT)

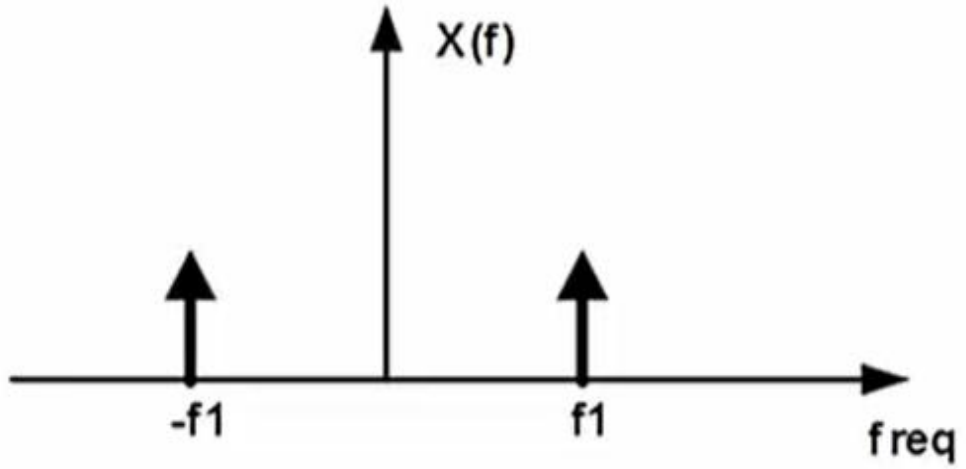
$$\hat{x}(k/N) = \sum_{n=0}^{N-1} x(n)\, e^{-i2\pi k n/N}$$

$$k = [0, N-1]$$

# Negative frequency, redundancy

All real signals have these negative frequency components. It is called the image.



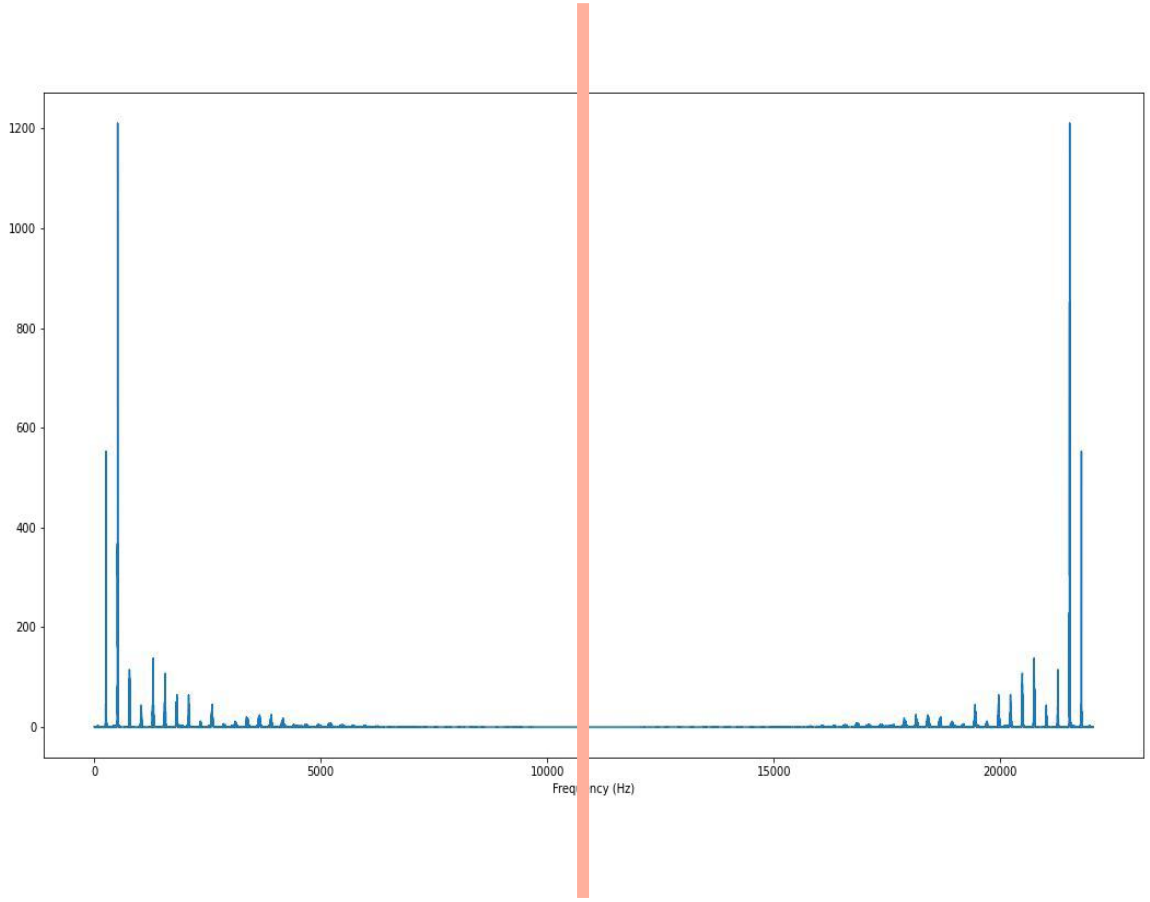

$$e^{j\omega t} = \cos(\omega t) + jsin(\omega t)$$

$$\cos(\omega t) = \frac{e^{j\omega t} + e^{-j\omega t}}{2},$$

$$\sin(\omega t) = \frac{e^{j\omega t} - e^{-j\omega t}}{2j}$$

A sinusoidal signal is often called a single-tone because its frequency domain representation contains only one frequency

# Redundancy in DFT




- $k = \frac{N}{2}, F\left(\frac{N}{2}\right) = s_r/2$
- The Nyquist Frequency

# From DFT to Fast Fourier Transform

- DFT is computationally expensive ($N^2$)
- FFT is more efficient ($Nlog_2N$)
- FFT exploits redundancies across sinusoids
- FFT works when N is a power of 2.

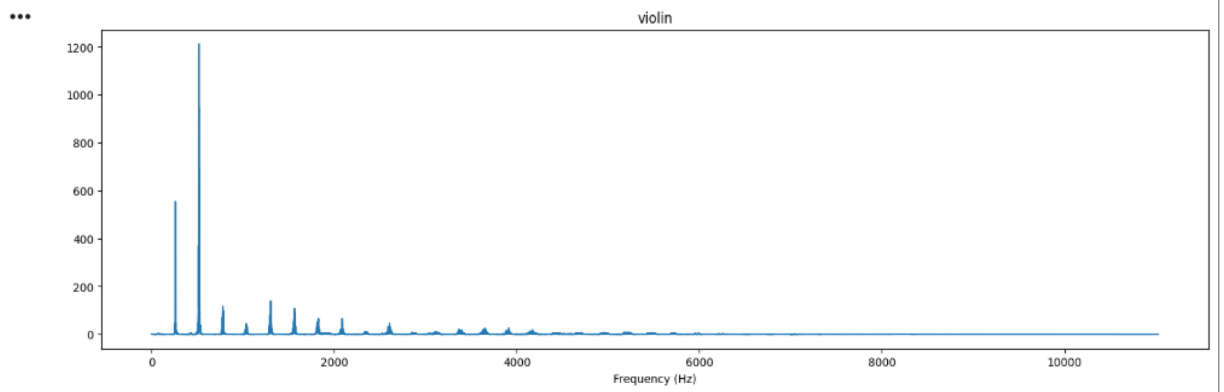
```
def plot_magnitude_spectrum(signal, sr, title, f_ratio=1):
    X = np.fft.fft(signal)
    X_mag = np.absolute(X)

    plt.figure(figsize=(18, 5))

    f = np.linspace(0, sr, len(X_mag))
    f_bins = int(len(X_mag)*f_ratio)

    plt.plot(f[:f_bins], X_mag[:f_bins])
    plt.xlabel('Frequency (Hz)')
    plt.title(title)
```

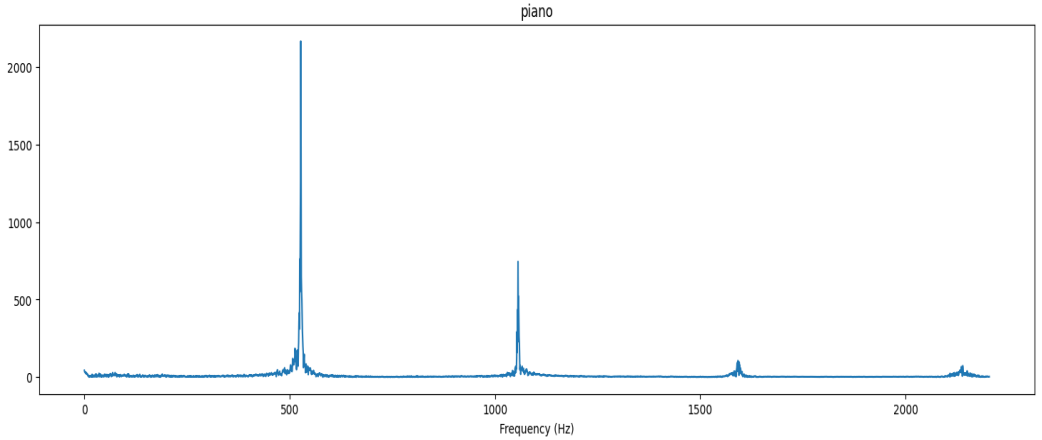
```
plot_magnitude_spectrum(violin_c4, sr, "violin", 0.5)
```



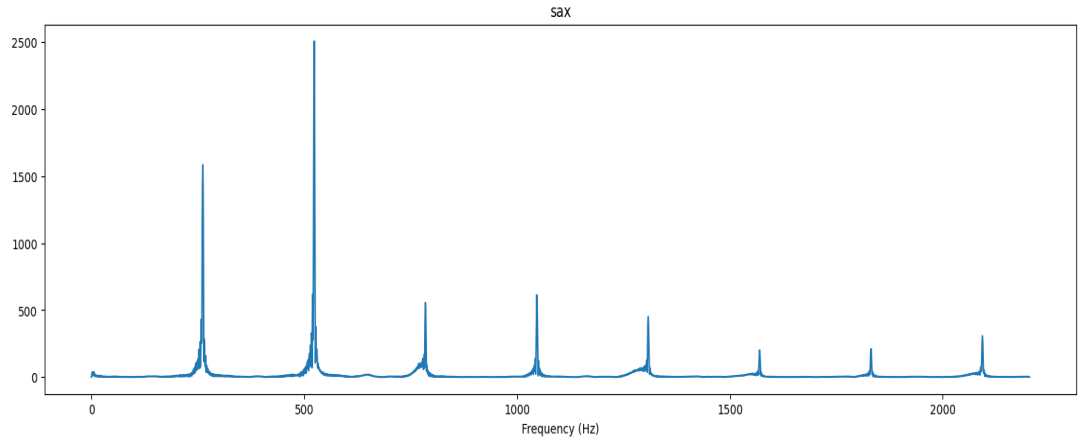

https://colab.research.google.com/drive/1E5uMNAa1mXX18-cN1bx9QifN6jxga2Nu#scrollTo=_kh1MK2Qi94q

# Magnitude spectrums

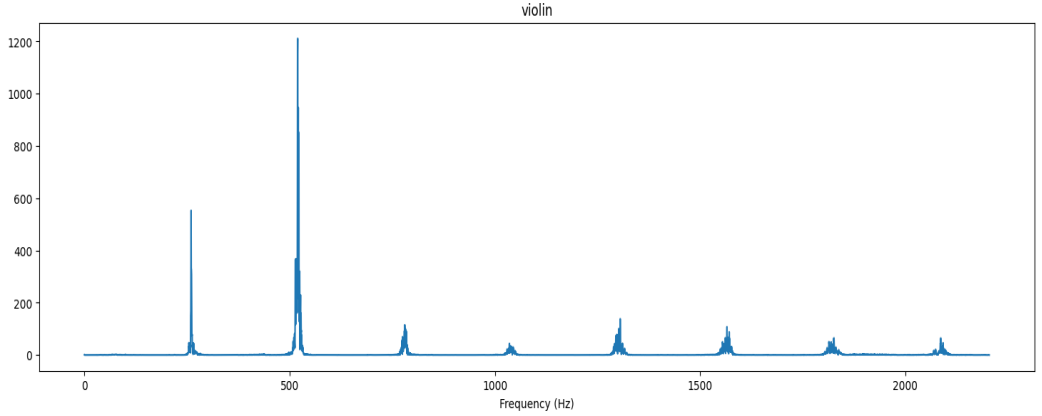
## Piano



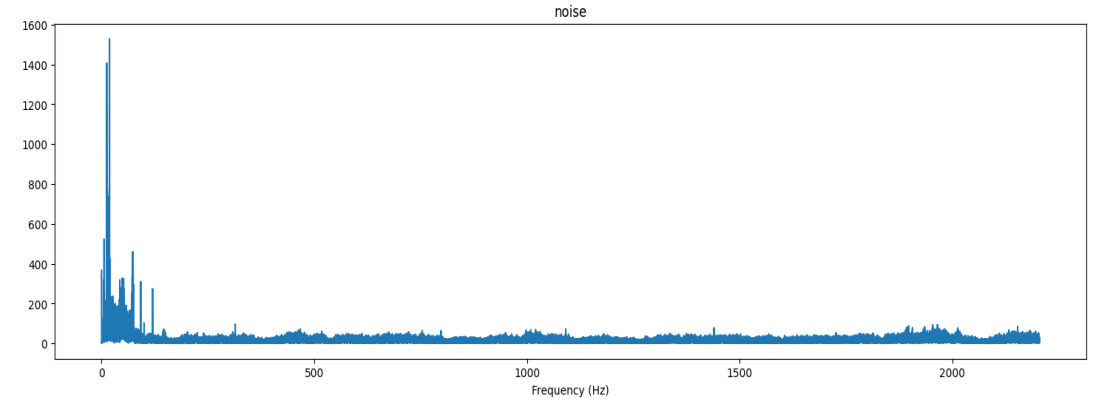

## Sax



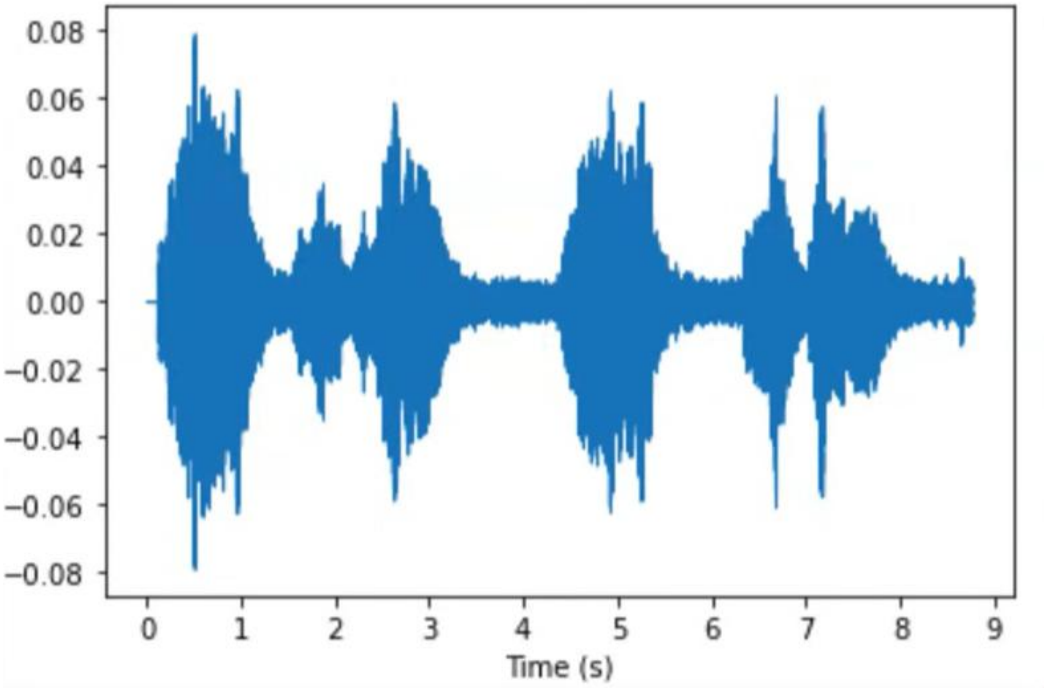

## Violin



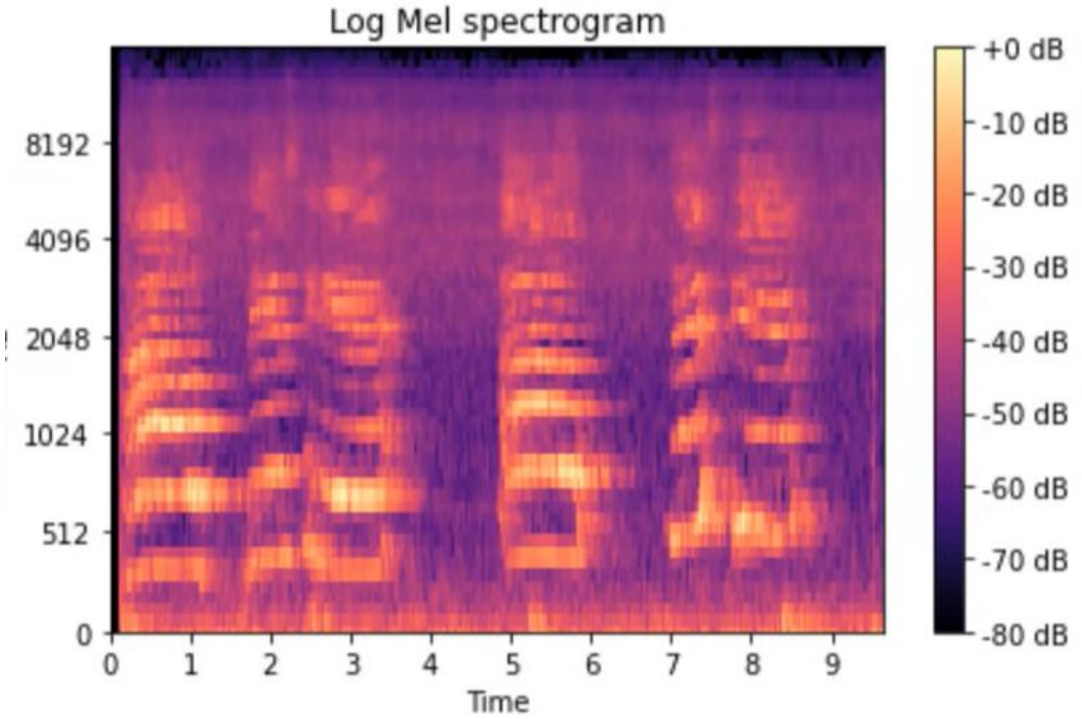

## Noise



# Time domain vs Frequency domain




Raw waveform: Time domain

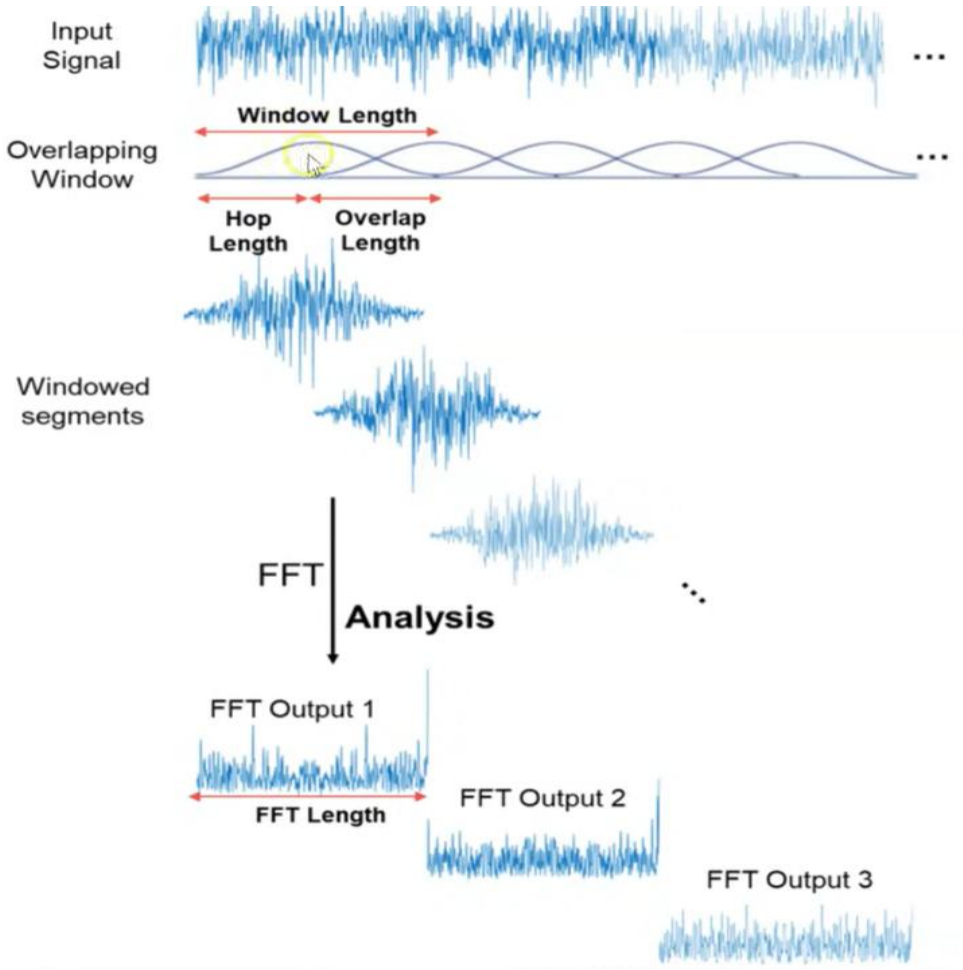



Log Mel Spectrogram: Time-Frequency domain, aperiodic signals

# What about aperiodic signals

Fourier Transform

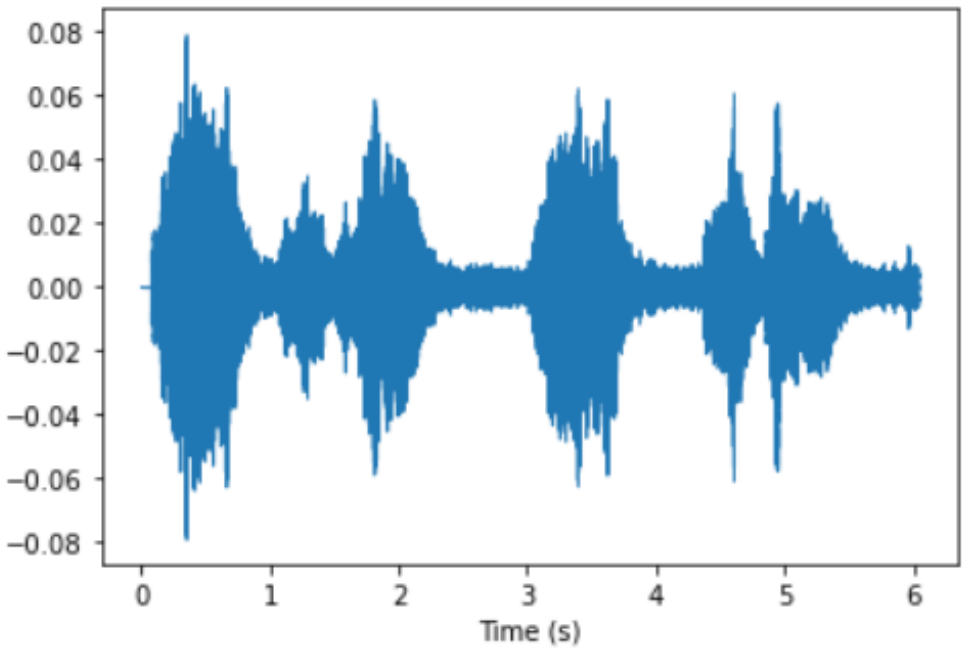
- Windowing
- Short-Time-Fourier Transform (STFT)
- Spectrogram



## Load sample audio file and display the raw waveform (time domain)

In [2]:
```
y, sr = librosa.load('h_1.wav', sr=32000)
librosa.display.waveshow(y, sr= sr, x_axis='s')
print("The sampled audio is returned as a numpy array (time series) and has ", y.shape, " number of samples")
print("The 10 randomly picked consequitive samples of the audio are: ", y[3000:3010])
```

```
The sampled audio is returned as a numpy array (time series) and has  (193195,)  number of samples
The 10 randomly picked consequitive samples of the audio are:  [-0.00938309 -0.01124619 -0.00986272 -0.00690551 -0.00444599 -
0.00405673
 -0.00585411 -0.00779615 -0.00679207 -0.00436785]
```




In [3]:
```
# Hear the audio
Audio('h_1.wav')
```

In [4]:

```
# Size of the Fast Fourier Transform (FFT), which will also be used as the window length
n_fft=1024

# Step or stride between windows. If the step is smaller than the window length, the windows will overlap
hop_length=320

# Specify the window type for FFT/STFT
window_type ='hann'

# Calculate the spectrogram as the square of the complex magnitude of the STFT
spectrogram_librosa = np.abs(librosa.stft(y, n_fft=n_fft, hop_length=hop_length, win_length=n_fft, window=window_type)) ** 2

print("The shape of spectrogram_librosa is: ", spectrogram_librosa.shape)
print("The size of the spectrogram is ([(frame_size/2) + 1 x number of frames])")
print("The frame size that we have specified is the number of samples to consider for the STFT. In our case, it is equal to
print("The number of frames depends on the total length of the sampled signal, the number of samples in each frame and the h
```

```
The shape of spectrogram_librosa is:  (513, 604)
The size of the spectrogram is ([(frame_size/2) + 1 x number of frames])
The frame size that we have specified is the number of samples to consider for the STFT. In our case, it is equal to the n_ff
t 1024  samples
The number of frames depends on the total length of the sampled signal, the number of samples in each frame and the hop lengt
h.
```

# Reference


- A. V. Oppenheim, and A. S. Willsky. Signals & Systems. Pearson Education, 2013. (Chapter 1-3, 5, 7)
- L. R. Rabiner, and R. W. Schafer, Introduction to Digital Speech Processing, Foundations and Trends in Signal Processing 1 (1-2), 1-194, 2007.