# ML for Audio Study Group

## Session 1:

December 14, 2021, 5 PM CET

hf.co/join/discord

**Omar Sanseviero**

**Vaibhav (VB) Srivastav**

JOIN

# Suggested readings before this session

- https://nbviewer.org/github/fastaudio/fastaudio/blob/master/docs/Introduction%20to%20Audio.ipynb
- SLP 26.1 to 26.5 https://web.stanford.edu/~jurafsky/slp3/

# Introduction

## Omar Sanseviero (https://twitter.com/osanseviero)

- ML Engineer at Hugging Face
- Previously
    - SWE at Google Assistant
    - Co-founder AI Learners

## Vaibhav Srivastav (https://twitter.com/reach_vb)

- MS student @ Uni Stuttgart/ Working Student @ Deloitte Tax
- Previously
    - Strategy @ Deloitte Consulting

# Organisation

- **Community-led!**
  - We'll kick off with some basics, but we'll decide collaboratively where we want to focus
  - Anyone can participate!
  - Members of the HF team and other cool collaborators will join.
- Expectation
  - Before each session: **Read/watch related resources**
  - During each session, you can
    - Ask question in the forum
    - Present  a short (~10-15mins) presentation on the topic (agree beforehand)
    - Participate a bit more passively (that's also ok and you're welcomed!)
  - Before/after:
    - Keep discussing/asking questions about the topic
    - Share interesting resources

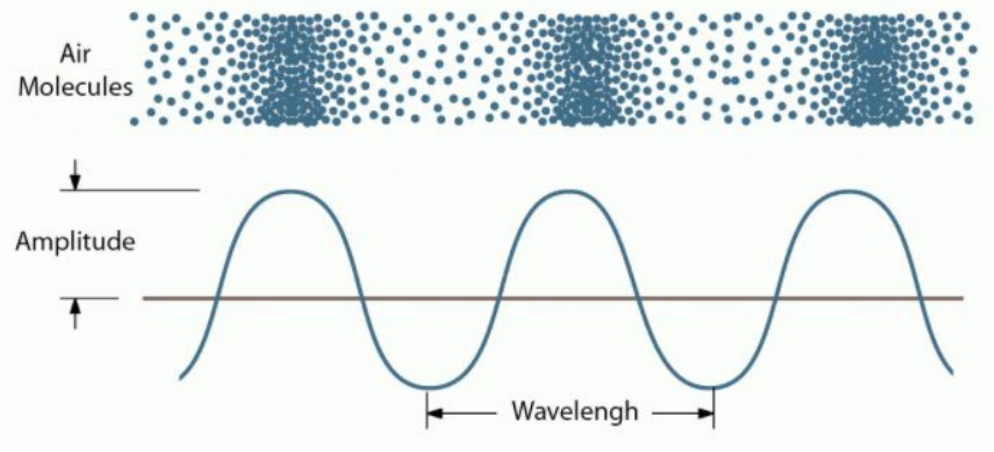
# Timeline

- Dec 14: Kick off session
- **Dec 21: ASR Deep Dive**
- Jan 4: TTS Deep Dive
- Jan 18 and forward:
  - Paper discussions
  - Invited speakers
  - Deep dive into a specific task

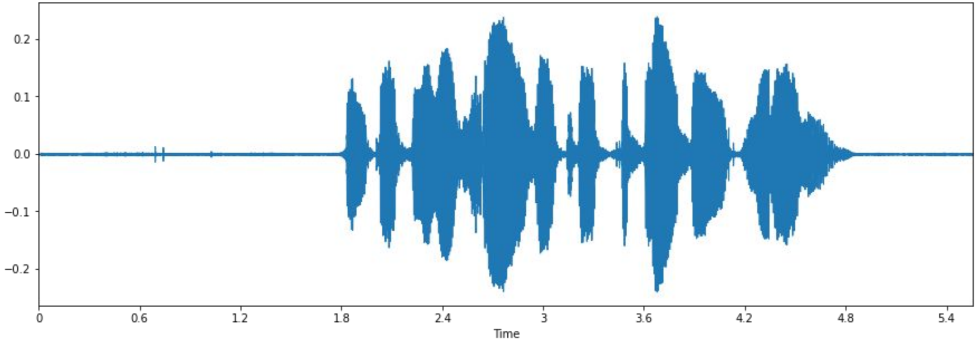
# Intro to Audio Data

# What is sound?

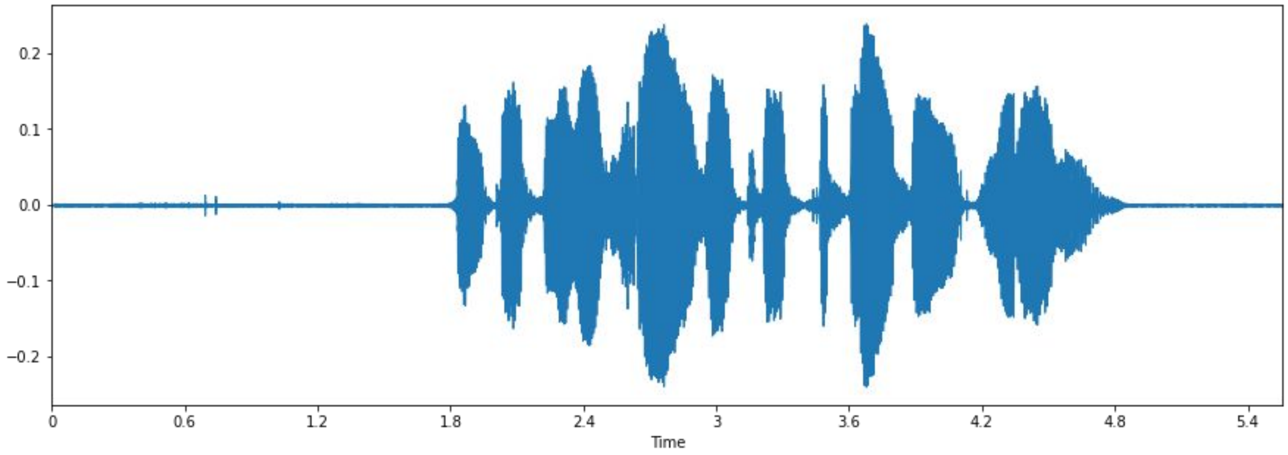
# What is sound?



# What is sound?

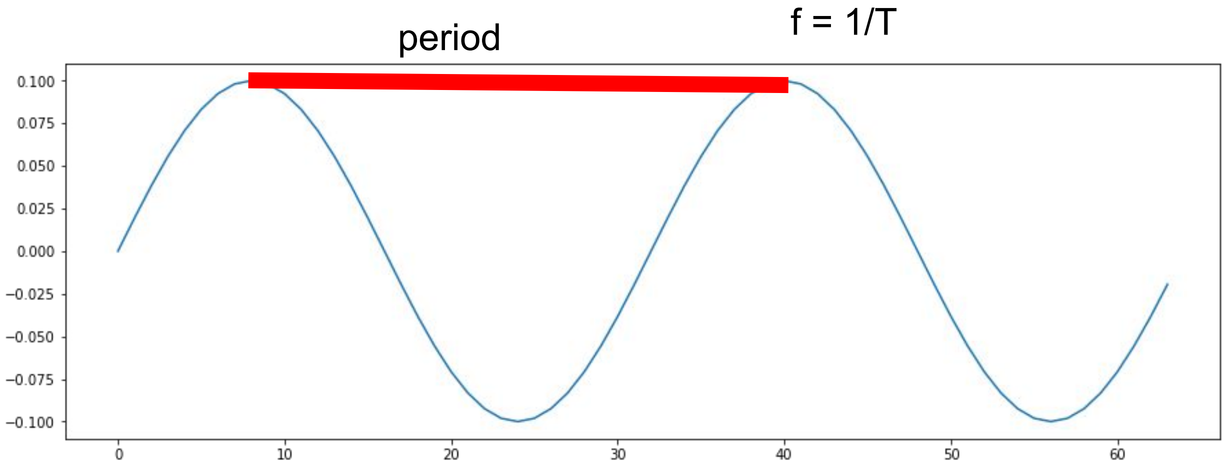


# Waveform



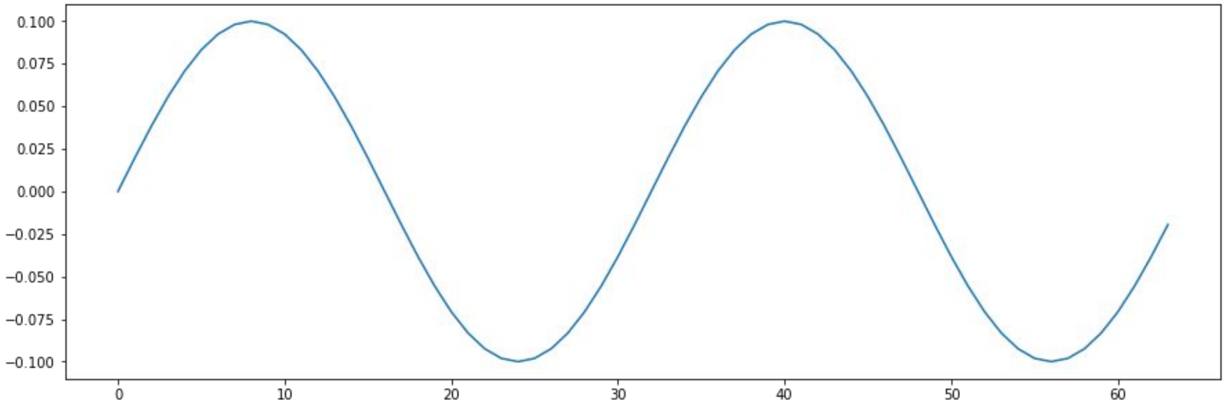
# Frequency

- Cycles per second of a wave



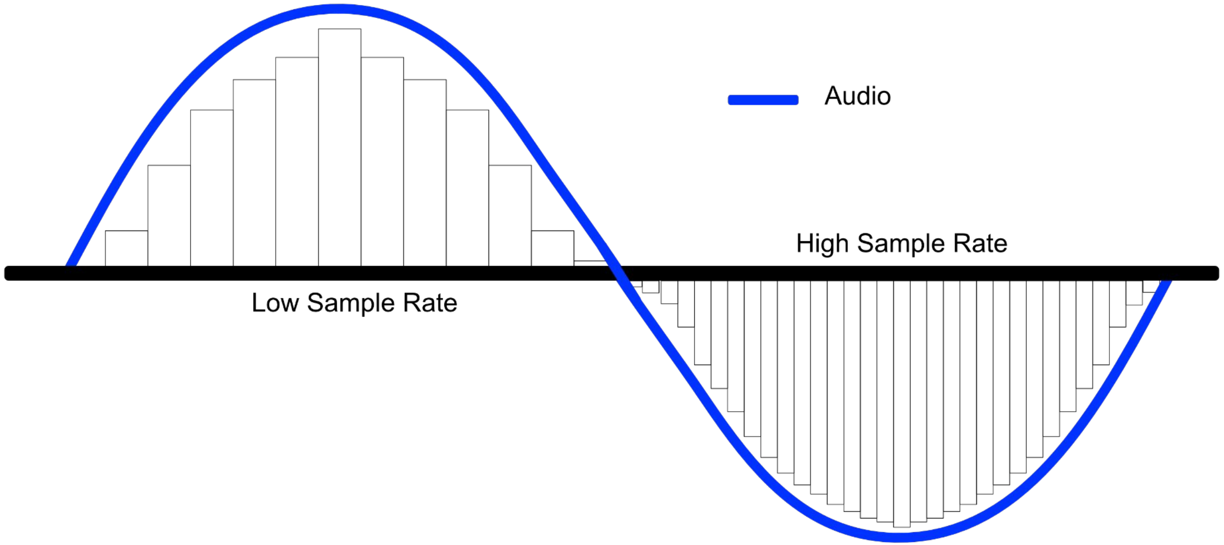
# Frequency

- Cycles per second of a wave
  - Human hearing ranges from 20hz to 20000hz
  - 500 hz = 500 cycles per second
  - 1 cycle = 16000/500 = 32 samples

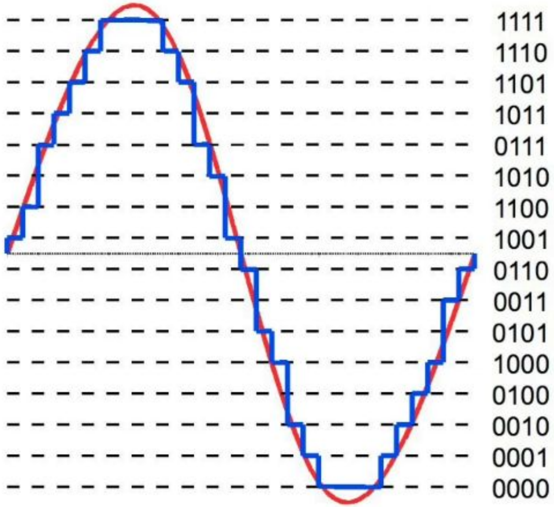


# Analog to digital conversion

- Sampling: sample at regular points in time
- Quantization: amplitude is represented in bits



# Quantization

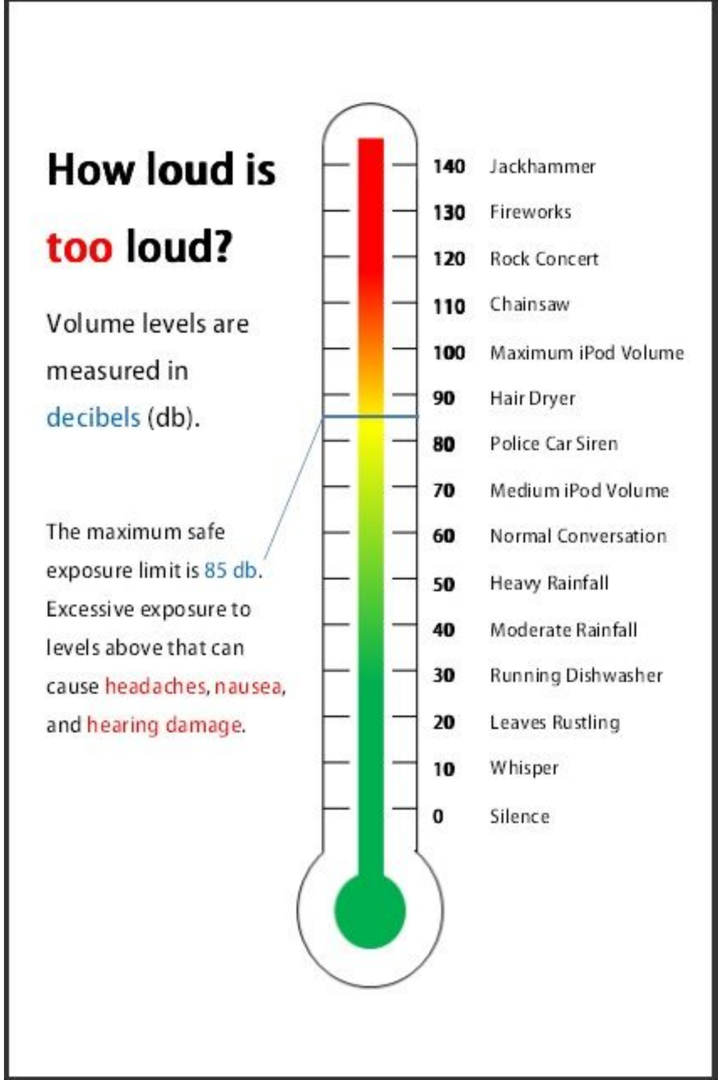


# Sampling rate

- Sample rate = 40,000 Hz
- Bit depth = 16 bits
- ((16*40,000)/(1,048,576*8))*60 = 4.58Mb of data for one minute of audio!

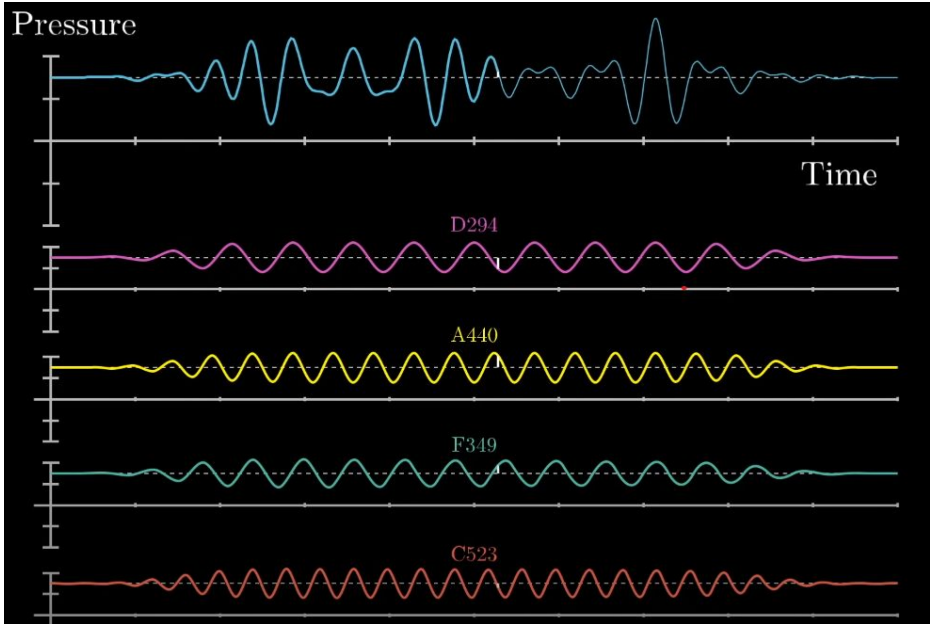
# Intensity and Loudness

- Intensity: rate at which energy is transferred
  - Measured in decibels
  - 10x increase in energy of wave results in a 10 dB increase of sound
- Human perception goes from 0dB to 100dB
  - 10,000,000,000x range
- Loudness: subjective perception
  - Depends on many factors (e.g. age)



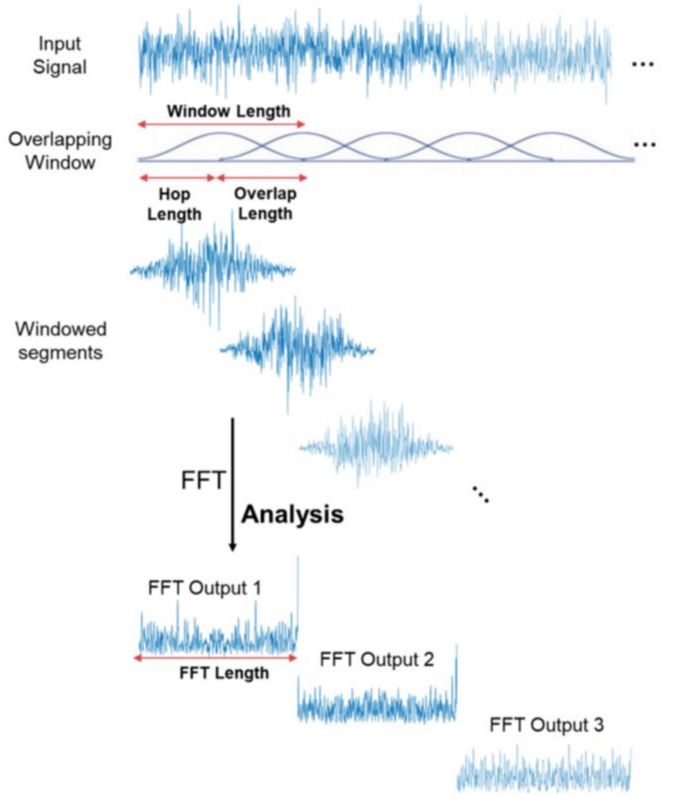
# Returning to sound data

- Many simple sounds
- What can we do with it?



# stft

- Many simple sounds
- What can we do with it?
  - We can decompose a signal into a set of waves with short-time Fourier transforms

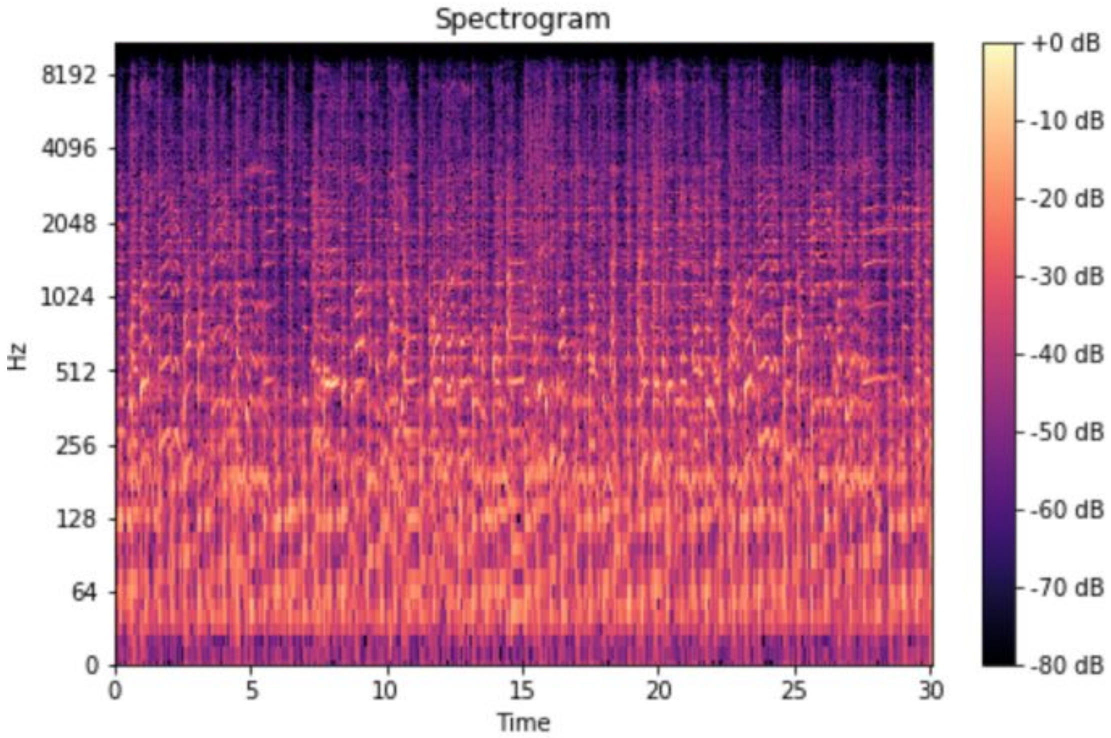



Source: https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53

# stft

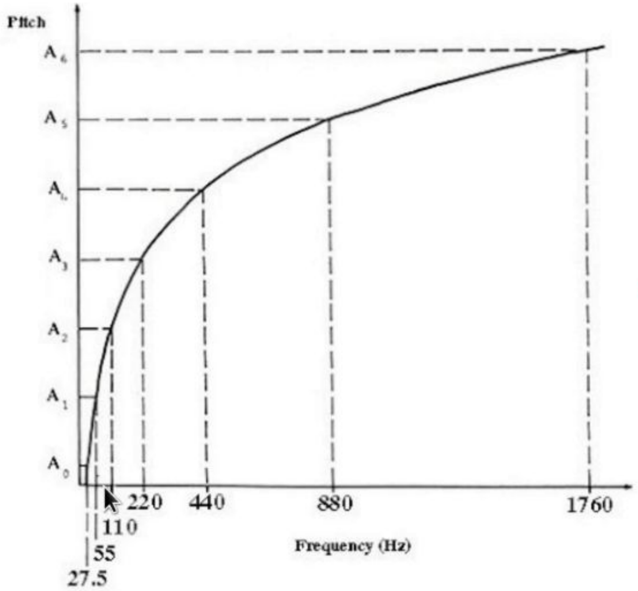
3 dimensions

- time
- frequency
- intensity



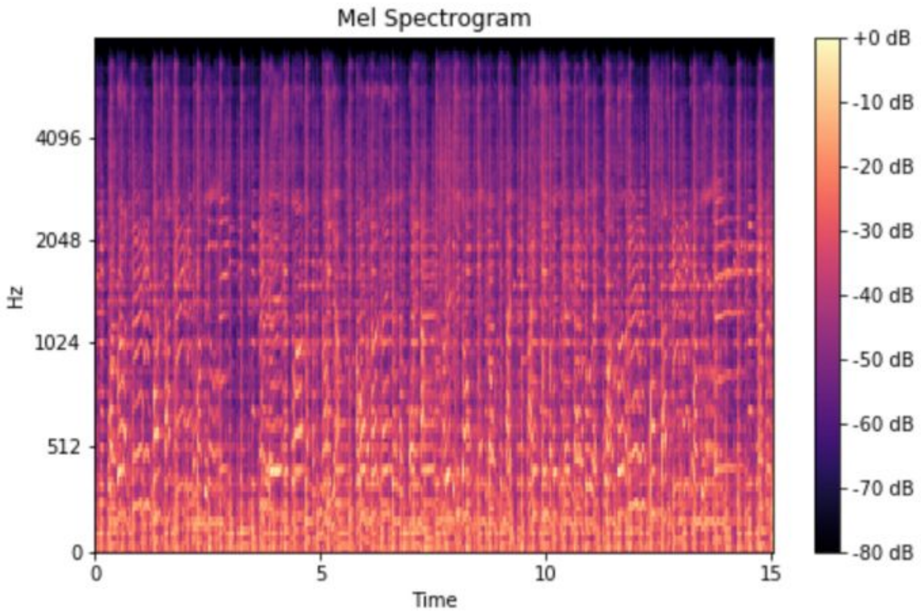
# Pitch

- Human perception of frequency
  - Logarithmic:
    - 100->200hz conveys as much info as 10K to 20K hz



# Pitch

- Human perception of frequency
  - Logarithmic:
    - 100->200hz conveys as much info as 10K to 20K hz
  - Unit: **Mel**
    - equal distances in pitch sounded equally distant to the listener



# Traditional ML approach vs DL approach

- Traditional
  - Compute manually features out of the spectrogram
    - Amplitude envelope
    - Band energy ratio…
  - Feed those features to a traditional ML model
- Deep Learning approach
  - Feed spectrogram directly

# Automatic Speech Recognition

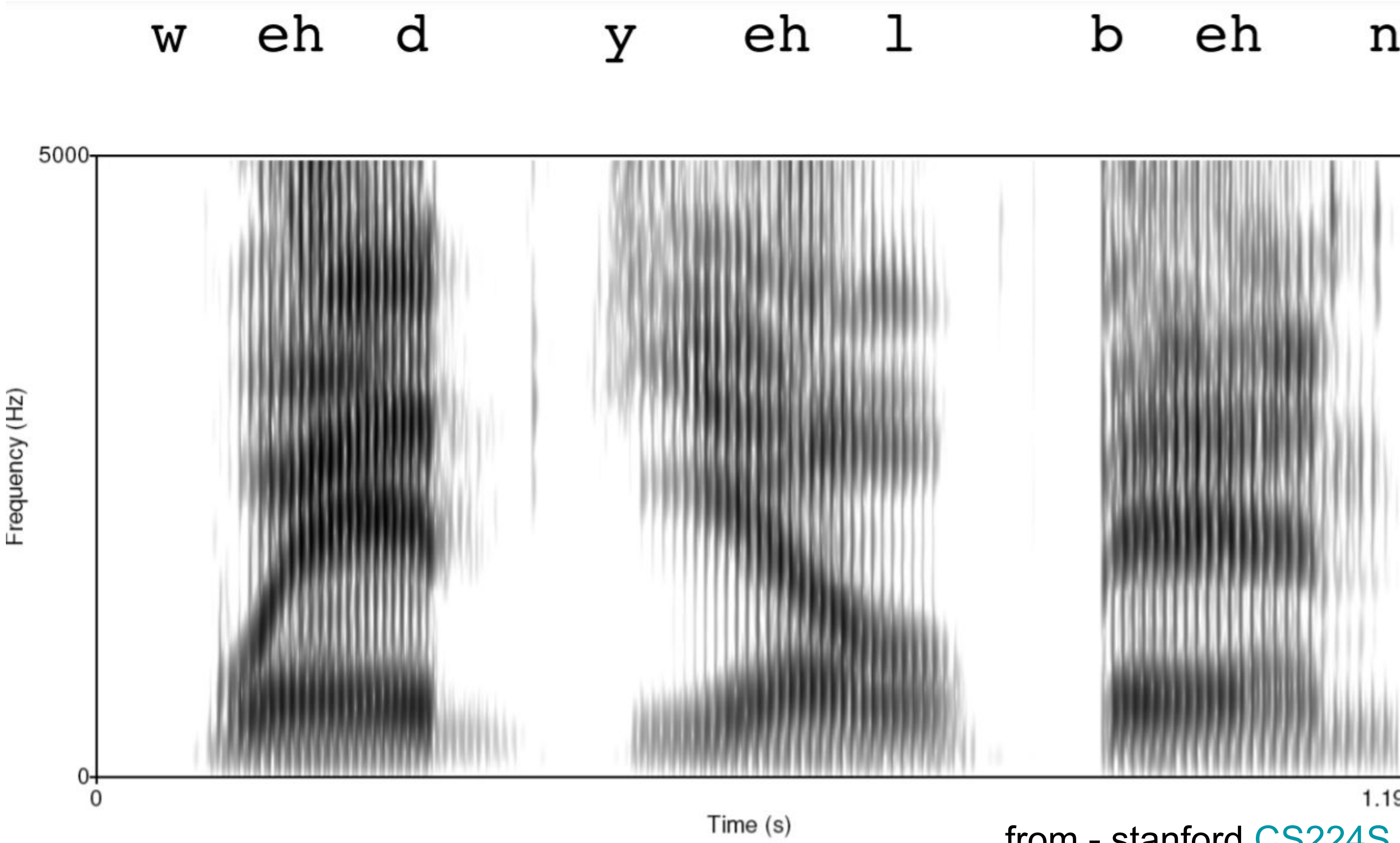
# What is ASR?

It’s time for lunch!

from - SLP3, Ch 26

# Why is ASR tough?

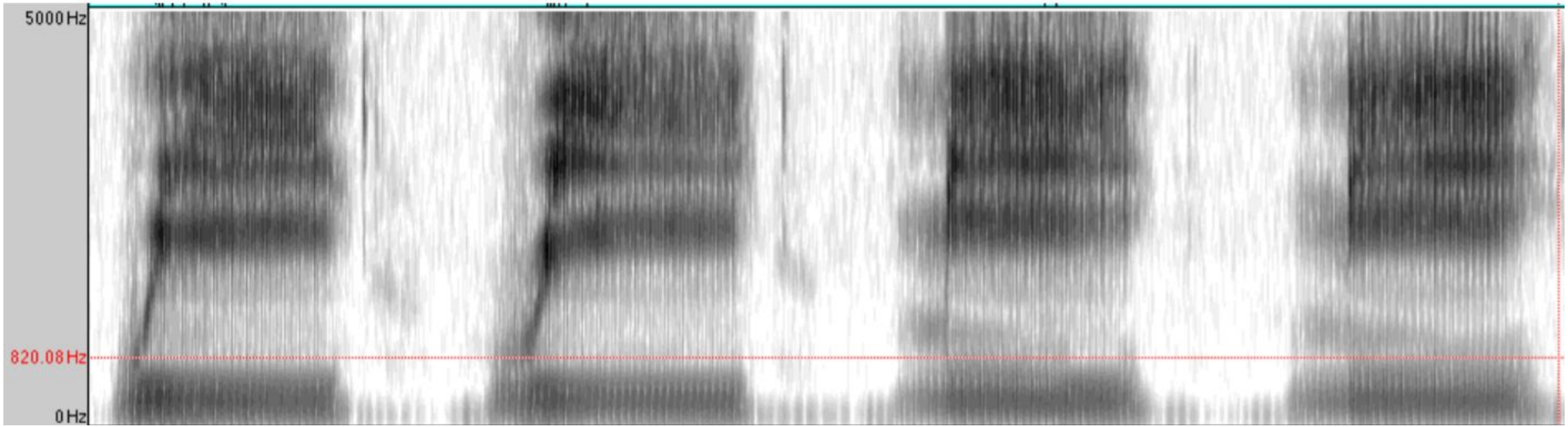
Different variations of "eh"




from - stanford [CS224S]

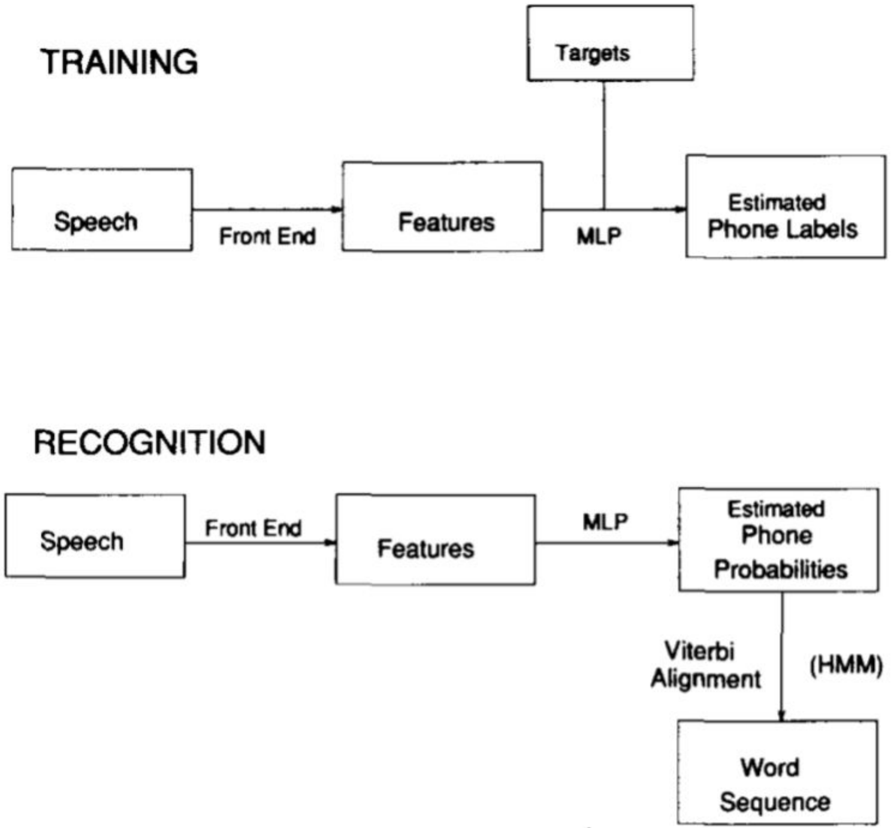
# Different variations of “iy” in context

w iy r iy m iy n iy



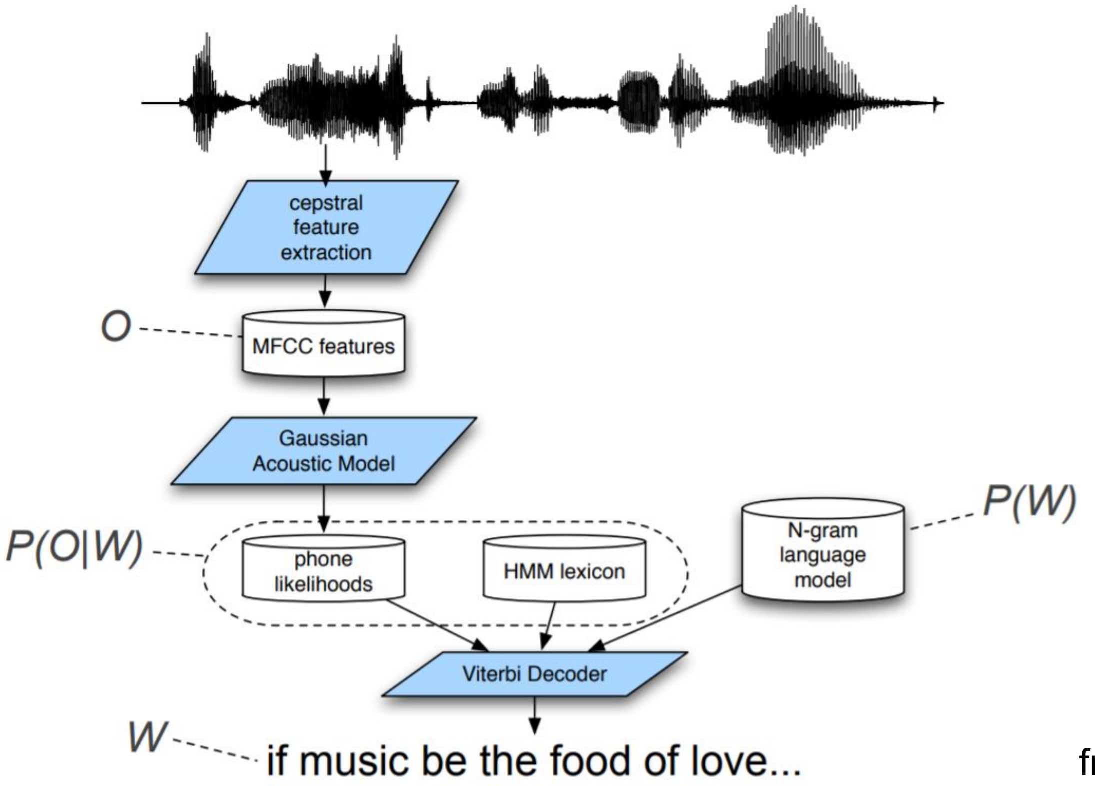

from - stanford CS224S

# Blast from the past



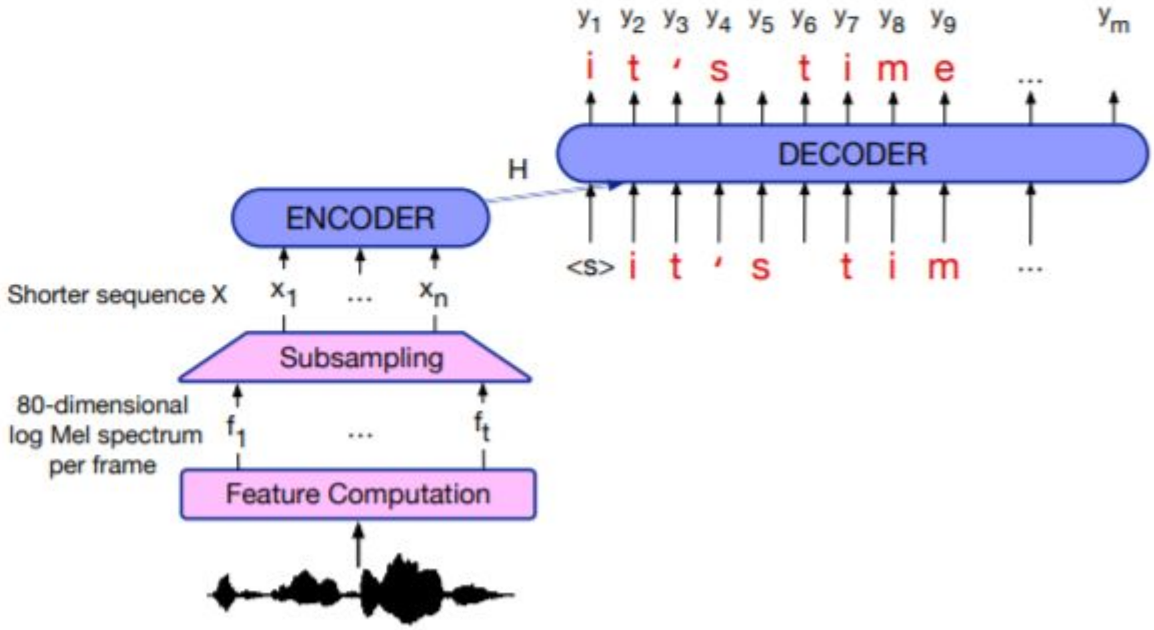

from - stanford [CS224S]

# Blast from the past




from - stanford CS224S

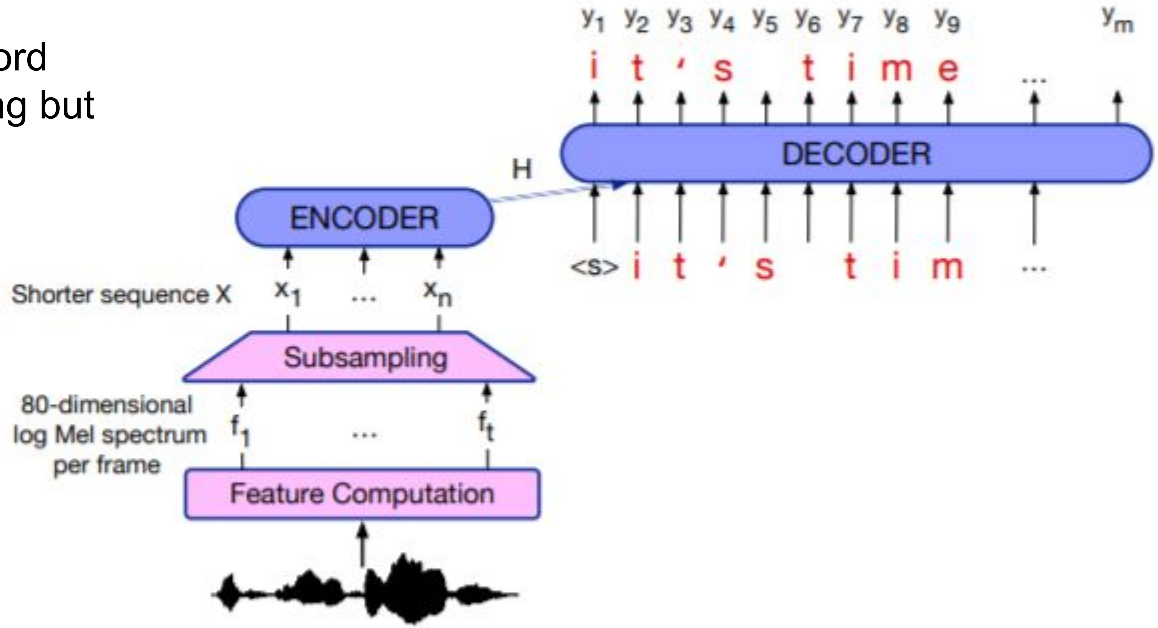
# Modern architecture for ASR




from - SLP3, Ch 26

# Modern architecture for ASR

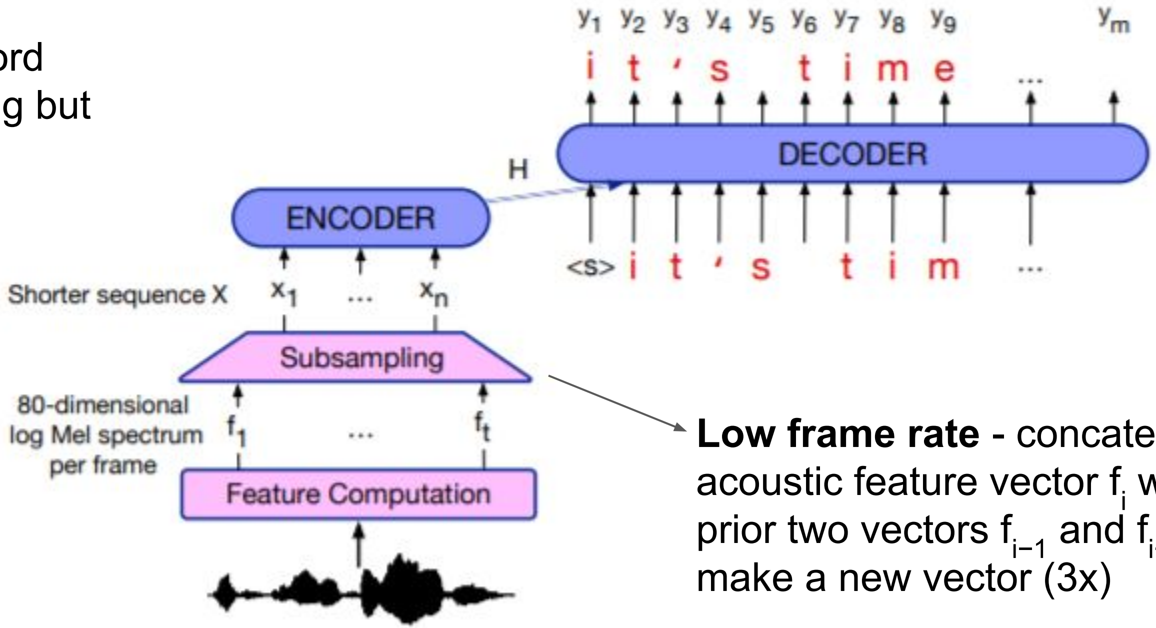
However, a single word might be 5 letters long but may take 2 seconds




from - [SLP3, Ch 26]

# Modern architecture for ASR

However, a single word might be 5 letters long but may take 2 seconds




**Low frame rate** - concatenate the acoustic feature vector $f_i$ with the prior two vectors $f_{i-1}$ and $f_{i-2}$ to make a new vector (3x)

from - [SLP3, Ch 26]

# Modern architecture for ASR

How do we exactly know which part of X (audio) maps to which part of Y (text)?

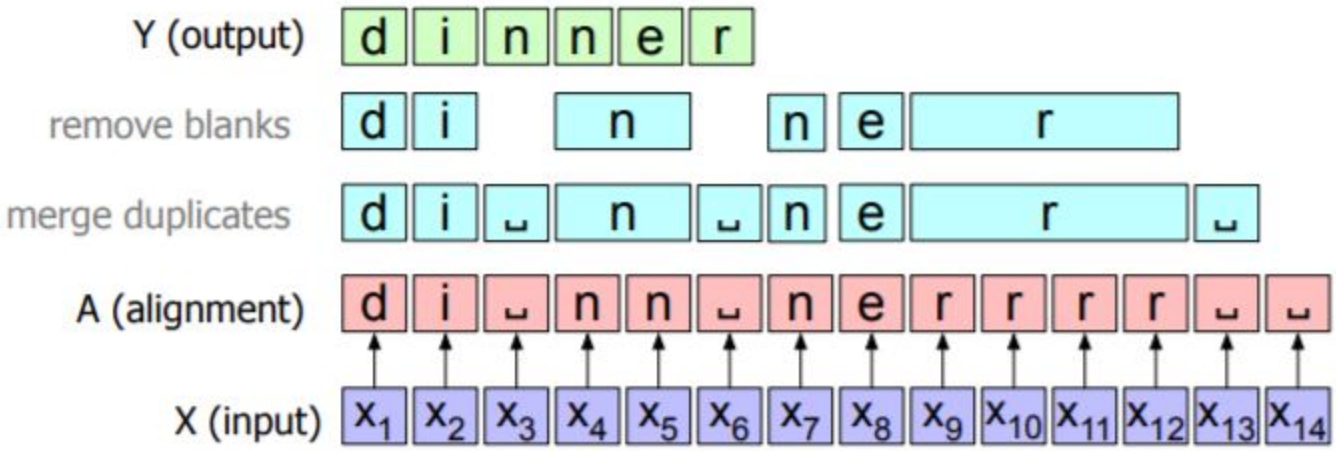
from - SLP3, Ch 26

# Connectionist Temporal Classification (CTC)

1. output a single character for every frame of the input
2. each input is mapped to an output
3. apply a collapsing function that combines identical letters
4. resulting in shorter output text sequence

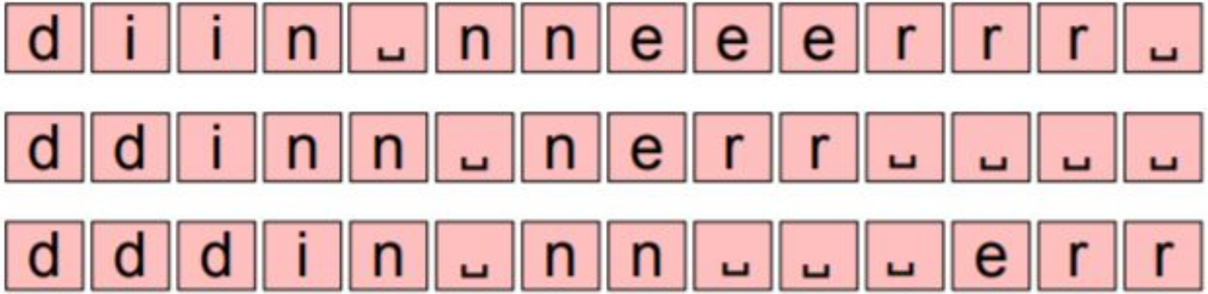
from - SLP3, Ch 26

# Connectionist Temporal Classification (CTC) | In action




from - [SLP3, Ch 26]

# Multiple alignments produce the same transcription




from - [SLP3, Ch 26]

# Current SoTA

1. Wav2Vec 2.0 - Convolutional transformer + masked audio modeling

2. Conformer - Convolutional augmented transformers (models both local and global dependencies)

3. ContextNet - CNN-RNN transducer network (introduces a squeeze-and-excitation layer)

from - SLP3, Ch 26

# Next steps

- Next week: 2 short (10-20min presentations + discussion
  - Presentation 1: TTS Deep Dive (Vaibhav Srivastav)
  - Presentation 2: Open slot (Post your ideas on the Discord channel)
- Recommended resources
  - Intro to Audio Notebook
  - ASR Notebook
  - SLP Chapter 26.6

Thanks for tuning in!