EE-336 Project Speech Recognition

Presented by Neeraj Pandey

Speech Recognition :

1. Isolated Recognition :Named Entity Recognition
2. Speech to Text Conversion

Is .wav file enough to get results?

No, we need signal encoding : converting .wav file to arrays

STFT (Short Time Fourier Transform):

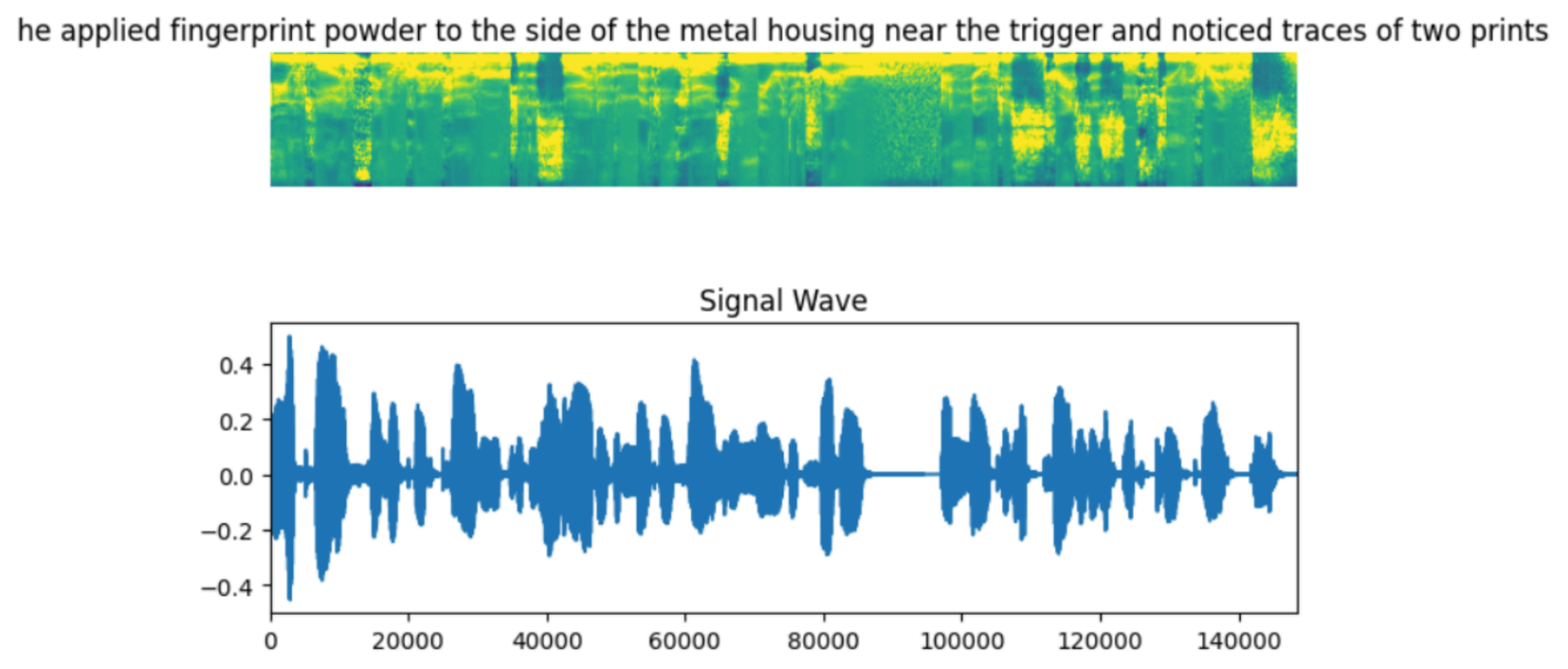
1. Widely used in speech signal processing.
2. Calculates FFT over a window of signal.
3. The project used Discrete Time STFT.

Mapping:

1. Vocabulary is created with all the characters in English vocab and punctuations
2. Each next character is probabilistically predicted.

Spectrograms:

1. Visual representation of associated frequencies.
2. STFT is normalized and plotted.

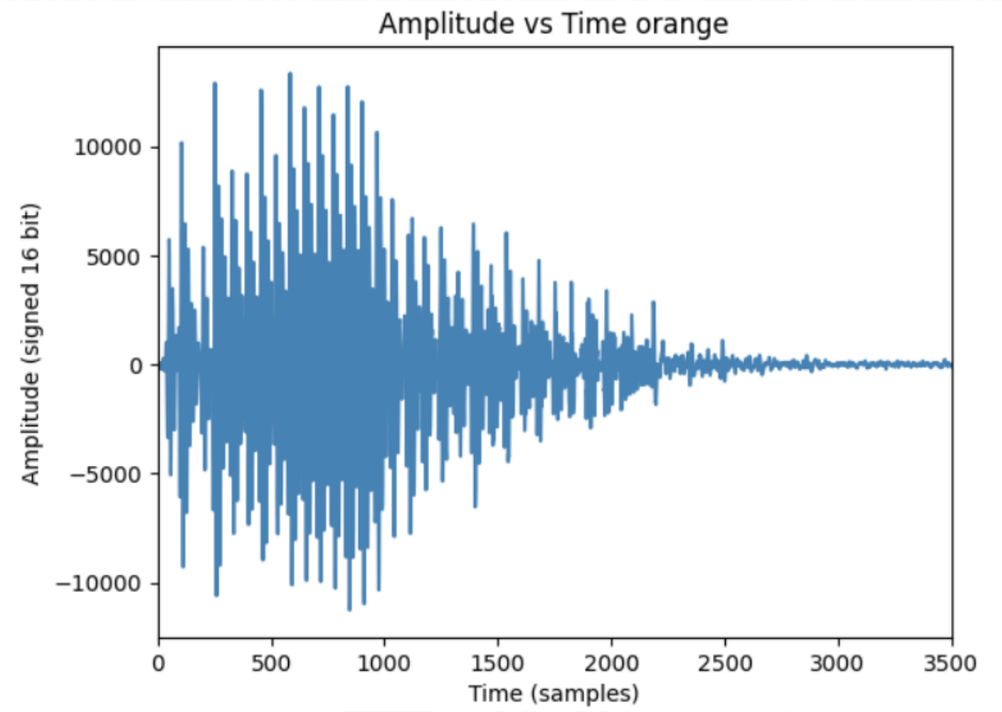


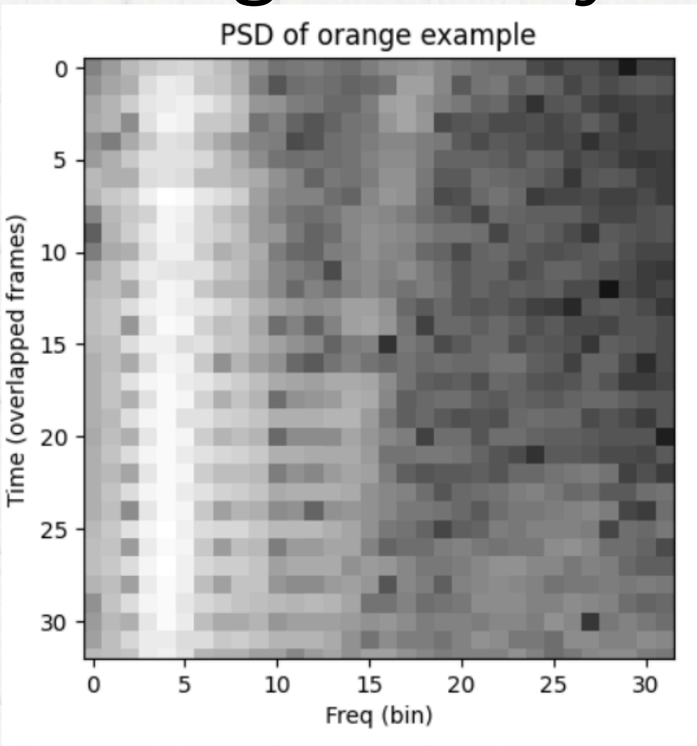
MFCCs(Mel Frequency Cepstral Coefficients):

1.MFCCs are derived from the STFT spectrogram.

2.MFCCs are a compact representation of the power spectrum of the signal.

3.Can be used as feature to speech recognition system.





1.Isolated Speech Recognition

We need to classify the sound to a class it belongs to (one vs all),
this task is also called as Named Entity Recognition.

Dataset:

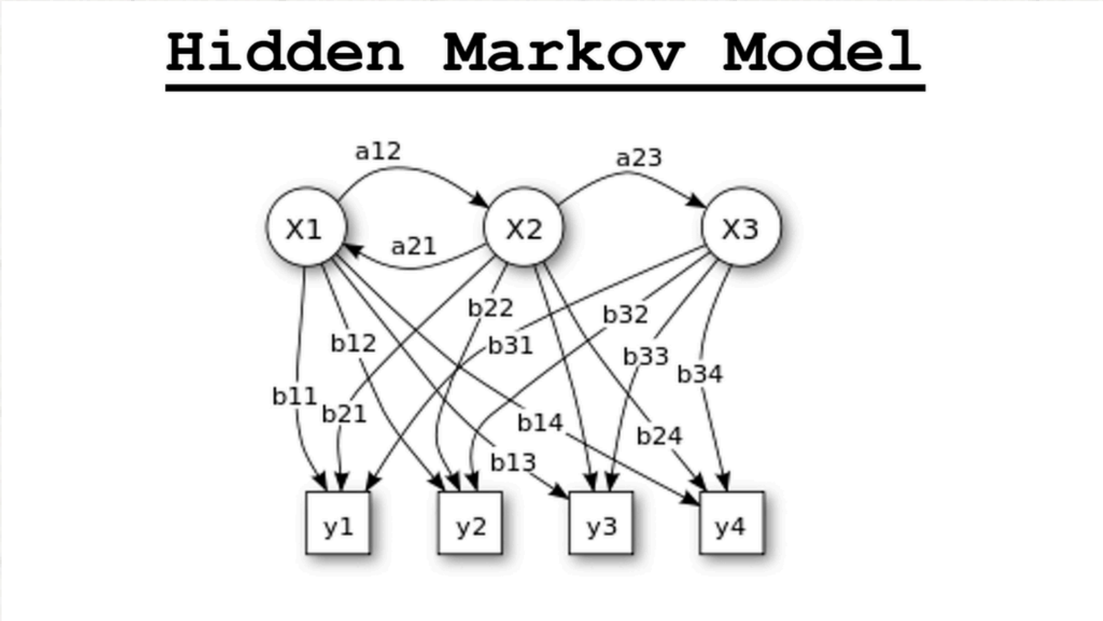
1. Dataset contains “.wav” files of saying one word only, viz.. “orange”, “apple”, “kiwi”, “god” etc. a total of 14 named entities.
2. Each word is mapped to an encoding for recognition.

Model Used: (Hidden Mrakov Model)

1. It’s a Markov Chain based approach with a set hidden states.
2. Hidden Markov Models (HMMs) involve a sequence of observable outcomes generated by an underlying sequence of hidden states.

1.Isolated Speech Recognition

3.Each state is associated with a probability distribution over observable outcomes.



Training and Results:

Trained a Custom Hidden Markov Model (gaussian) with 6 hidden states on a datase of 202 samples, achieved accuracy of 56.52% on test set(23 samples).

2. Speech to Text Conversion

Dataset:

1. A CSV file containing “.wav” file name and corresponding transcription.
2. A folder containing “.wav” files corresponding to the file name given.

Model Used: (CNN and RNN based)

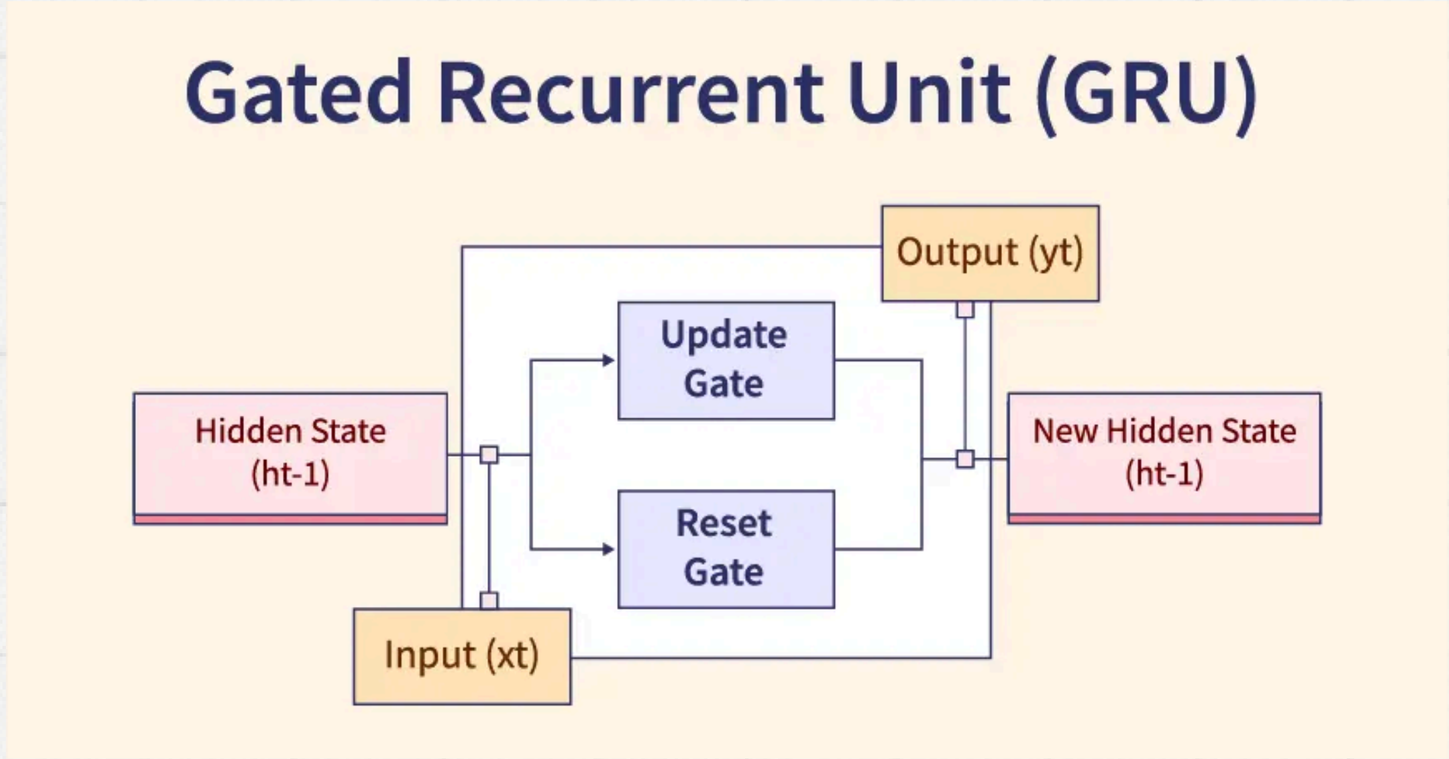
1. The model uses Convolutional Neural Networks stacked with Recurrent Neural Netwroks.
2. Algorithm uses Adam optimizer (Back Propagation with normalized weights).

Recurrent Neural Networks:

1. RNNs apply weights recursively, introducing variability in predictions based on input sequence context.
2. RNNs maintain evolving internal state, influencing predictions non-deterministically over time

2. Speech to Text Conversion

3. RNNs when used single layered, can be used deterministically, here, its used semi stochastically/stochastically.
4. We've used GRU(Gated Recurrent Unit) version of RNNs.

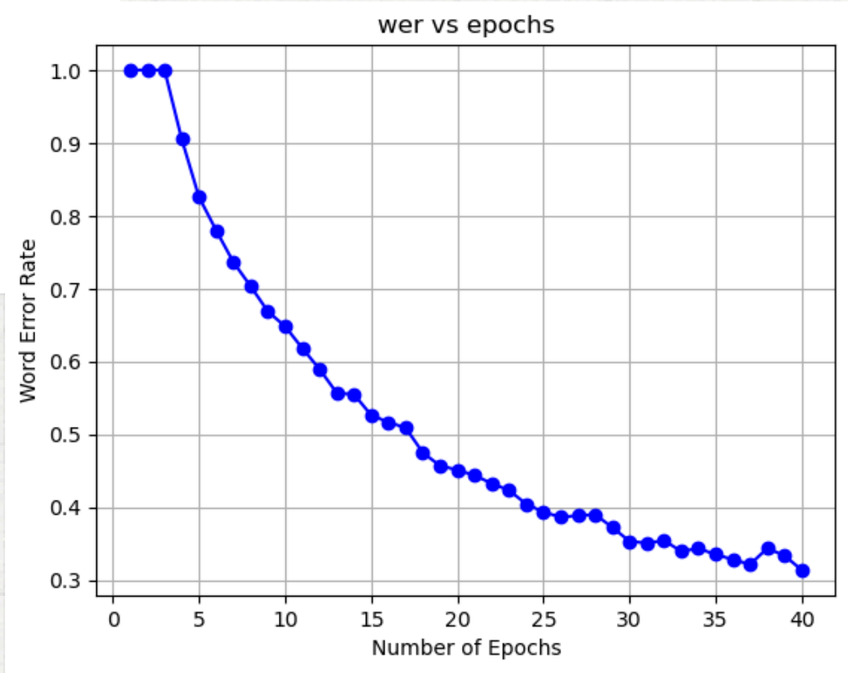


2. Speech to Text Conversion

Training and Results:

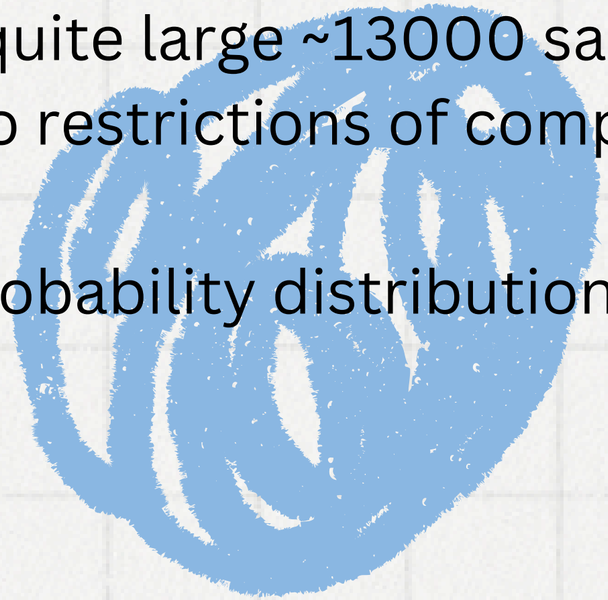
Trained 40 epochs with 8515 (6386+2129) samples using CTC(Connectionist Temporal Classification) loss. Achieved a word error rate of 0.3131 on test set.

```
Word Error rate : 0.3131
-----------------------------------------------------------------------------------------------------------
Target : with the means of earning an honest livelihood if so disposed
Prediction : with the means of arning an oneslively whodt if sodisposed
-----------------------------------------------------------------------------------------------------------
Target : six projects will be allocated to localities or relief areas in relation to the number of workers on relief rolls in t
hose areas
Prediction : six projects will be alicated to localities o relief areous in relation to the number of workers on relief roles i
n those arioas
-----------------------------------------------------------------------------------------------------------
Target : government is to be rendered impotent
Prediction : goveredment is to be renderd impitent
-----------------------------------------------------------------------------------------------------------
Target : having mentioned calcraft's name
Prediction : havingmentianed calcraf's name
-----------------------------------------------------------------------------------------------------------
Target : were found in the butler's pantry used by courvoisier
Prediction : were found and the butlerce pantry use by crvisie
-----------------------------------------------------------------------------------------------------------
```



Improvements:

1. Dataset for Speech to Text Conversion was quite large ~13000 samples, but was trained on 65% of original dataset due to restrictions of computational power.
2. Hidden Markov Models with several other probability distributions can be used for Named Entity Recognition.



References:

1. A tutorial on HMM and selected application in speech recognition. *By-Lawrence R. Rabiner, Fellow, IEEE.*
2. IMPLEMENTATION OF SPEECH TO TEXT CONVERSION USING HIDDEN MARKOV MODEL, *Published in 6th ICECA, 2022 ,by-Saveetha engg. College, Chennai.*

Thank you very much!