# RESEARCH AND DEVELOPMENT CS402

# Run-time conversion of speech from one language to another using deep learning

Mentors: Prof. Prosenjit Gupta
Mr. Vikas Malviya

# GROUP MEMBERS

| | |
|---|---|
| Deeptonabho Dutta | U101116FCS030 |
| Gaurav Mundhra | U101116FCS037 |
| K.N Raviteja | U101116FCS056 |
| Sabyasachi Mishra | U101116FCS104 |
| Shubhangi | U101116FCS127 |

# CONTENTS

# Objectives of the Study

*Objective 1:* To implement the project using machine learning libraries

*Objective 2:* To switch from machine learning libraries to deep learning libraries and neural nets

**Speech Recognition:**
Deepspeech and Tensorflow (or) HMM model

**Speech Translation:**
Tensorflow, NLTK and Keras

**Speech Synthesis:**
Tensorflow ,Matplotlib and Tensorboard

# Rationale of the Study

Language is a genuine problem when difficult to understand. Through our research, we aim to considerable solve the following problems:

- **Lingual Barrier -** To make language easy to understand and also to widen our knowledge in the fields of machine learning and deep learning.
- **Conferences** - To make speeches in international conferences accessible to all the dignitaries.
- **Education** - To instantaneously translate a teacher's lecture to different languages in a classroom.
- **Tourism** – To enable a tourist to communicate effectively with people in a new country.
- **Blind People** - To translate on-screen text to a blind person's native language.

# Rationale of the Study (Continued)

**Why not use previous models**

- **Pre-1970**

1. Speech recognition was implemented for less vocabulary of words(16-200 words).
2. User had to give a pause after every word for the system to recognize.

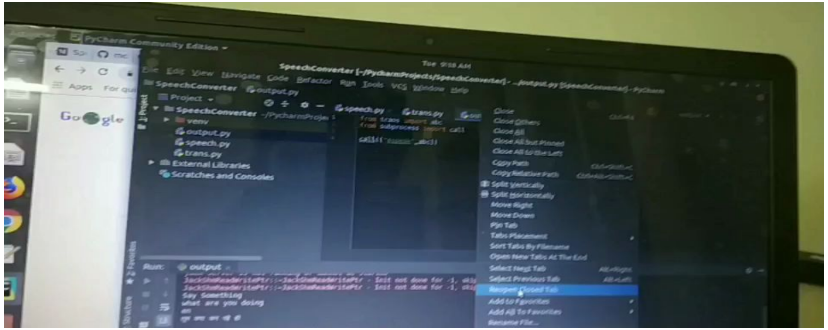
- **1970-2000**

1. Introduction of HMM model.
2. 30 seconds of speech required 100 minutes to be translated into another language.

- **2000s**

1. Introduction of LSTM which outperforms HMM models.
2. There was a increase in the performance and decrease in word error

# Results of the work done: Machine learning




**Successfully implemented with ML libraries and noted the Shortcomings -**

- Output is not that accurate
- Bad quality microphone is prone to noise
- Not real time

# Results of the work done: Deep learning (Continued)

```
| Adam | epoch: 003 | loss: 0.20745 - acc: 0.9607 -- iter: 09792/
Training Step: 468  | total loss: 0.20083 | time: 0.928s
| Adam | epoch: 003 | loss: 0.20083 - acc: 0.9631 -- iter: 09856/
Training Step: 469  | total loss: 0.19530 | time: 0.934s
| Adam | epoch: 003 | loss: 0.19530 - acc: 0.9668 -- iter: 09920/
Training Step: 470  | total loss: 0.19072 | time: 0.939s
| Adam | epoch: 003 | loss: 0.19072 - acc: 0.9670 -- iter: 09984/
Training Step: 471  | total loss: 0.19505 | time: 0.945s
| Adam | epoch: 003 | loss: 0.19505 - acc: 0.9640 -- iter: 10000/
--
predicted digit for 9_Alex_120.wav : result = 9
नौ
```

# Results of the work done: Deep learning

**Speech Recognition**

- In this approach  we build an LSTM recurrent neural network using the TFLearn high level Tensorflow-based library to train on a labeled dataset of spoken digits. Then we test it on spoken digits
- We label the Dataset with the help of the  extracted features (Frequency and width) from the .wav file and perform classification for phoneme
- This module is almost complete as there are few errors while training the model due to lack of high end GPU (RTX 1080 8 gb)

# Results of the work done: Deep learning

**Speech Synthesis**

- We use a TTS engine
- In this approach we convert raw text containing symbols like numbers and abbreviations into the  equivalent of written-out words
- we  assigns phonetic transcriptions to each word .
- Converting  the symbolic linguistic representation into sound .
- Yet to implement

# Results of the work done: Deep learning

**Speech Translation :**

We use Neural MachineTranslation:

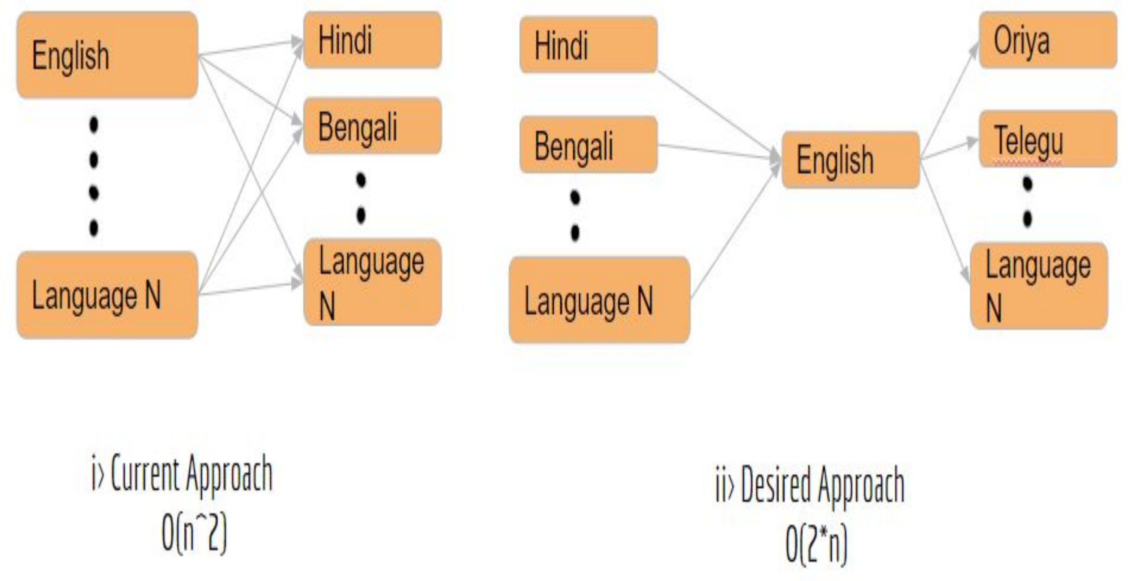
- RNN Encoder–Decoder : It consists of two recurrent neural networks (RNN) that act as an encoder and a decoder pair.
- The encoder maps a variable-length source sequence to a fixed-length vector, and the decoder maps the vector representation back to a variable-length target sequence.

# Future work

- Mastering the skills required in order to overcome the errors faced.
- Implementing the other two phases(Translation, Synthesis) using deep learning with large corpus and high-end GPU.
- Modifying the model according to our novelty for better performance.

# Novelty

- We have decided to train our model with human voice signal so that the synthesis becomes more real and with emotions.
- We plan on using English as an intermediate language for optimizing the complexity from O(n^2) to O(2*n)



# Difficulties faced

- Getting the required Dataset.
- Need of high-end GPU and good quality microphone.
- Problem using google colab.

# References


- Alex Waibel,Gate Alon Lavie, Lori Levin, Kevin Lenzo, Laura Mayfield Tomokiyo Juergen Reichert, “two-way sequence to sequence translator”, NAACL-Demonstrations '03 Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Demonstrations - Volume 4,Pages 29-30,Edmonton, Canada — May 27 - June 01, 2003.

- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation”, Google ,September 26, 2016 Problem using google colab.

- Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, Yoshua Bengio, “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation”, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, October 2014

- Li Deng ,Xiao Li,“Machine Learning Paradigms for Speech Recognition: An Overview”, IEEE Transactions on Audio Speech and Language Processing ,Volume 21(5): Pages -1060-1089 ,May 2013