# Building an ASR system for Indic Languages - Indian English

Mirishkar S Ganesh, IIIT Hyderabad (Mentor)

Siddharth Gupta, IIT Delhi
Rohit Reddy, Amrita Vishwa Vidyapeetham
Deepu C, ICFOSS

# Flow of Presentation

- **Problem Statement: Challenge with Indian English transcription**
- **Dataset Description**
- **ASR Pipeline**
- **Experiments**
- **Results Comparison**
- **Future Work**

# Problem Statement: Challenge with Indian English transcription

**Native language influence**

- L2 English speakers may have a huge L1 language influence in their English speaking language. For eg. If a person A, whose L1 is Malayalam, and there is another person B whose L1 is Telugu then the accent produced while they speak English could be completely different due to their L1 influence.

**Problem Statement**

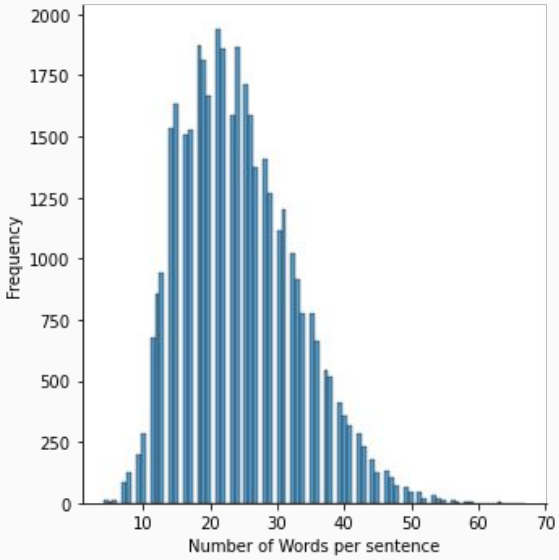
- In this project, an Indian English ASR system based on Hidden Markov Models (HMM) has been designed using Kaldi(Povey et al., 2011). We used available continuous English speech transcribed data obtained from non-native Indian English speakers in order to build an ASR system robust to across accent lines.

# Dataset Description(1/2)

- NPTEL lecture videos delivered in English.
- Consists of 39341 .wav audio files.
- Files are 16-khz, mono channel.
- Each file being considered as having one utterance.

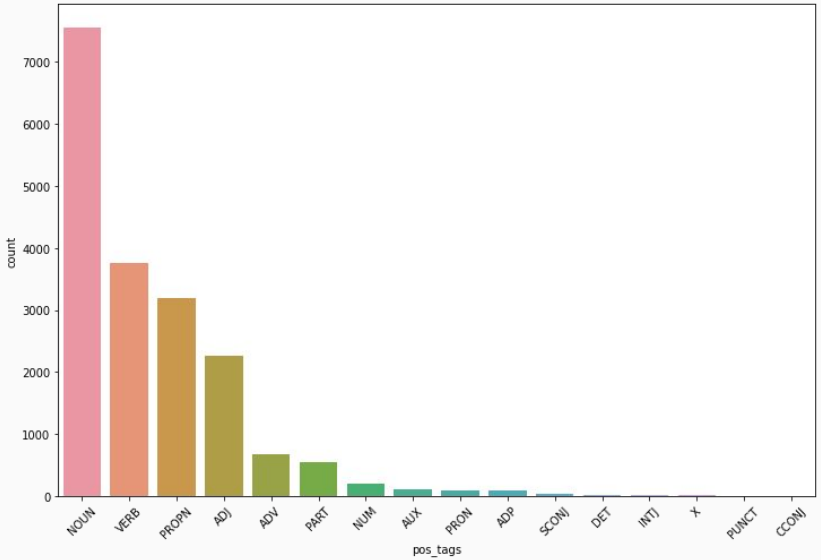
| | |
|---|---|
| count | 39341.000000 |
| mean | 24.184388 |
| std | 8.520111 |
| min | 4.000000 |
| 25% | 18.000000 |
| 50% | 23.000000 |
| 75% | 30.000000 |
| max | 67.000000 |

Per utterance word count

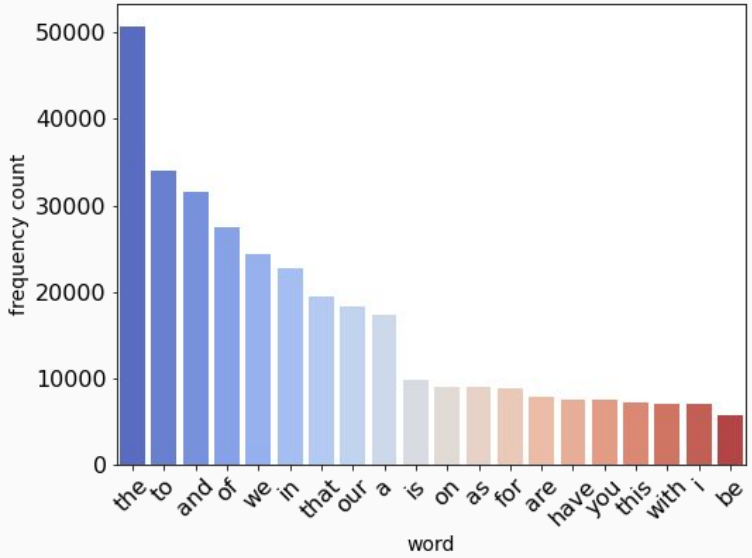



Frequency distribution for words per occurrence

# Dataset Description(2/2)




Frequency distribution of Part of Speech




Frequency distribution of vocabulary in corpus

# ASR Pipeline

**Text preprocessing:** lowercasing, substituting punctuations with space.

**Lexicon generation:** Through g2p library(Park, 2019). For eg. the phoneme for ‘thank you’ would be ‘TH AE NG K Y UW’.

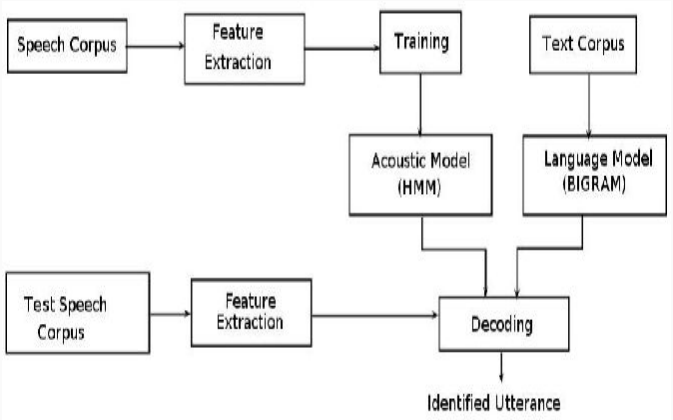
**Train/Test split:** An 80:20 split was performed on dataset.

**Feature extract:** MFCC features extracted from wav files, CMVN applied to perform normalization.

**Training:** HMM modeling is performed on normalized MFCCs.

**Language Model:** n-gram (n=1, 2, 4) using SRILM (Stolcke, 2002).

**Decoding:** Calculates likelihood of words forming a sequence, **finally utterance.**




Block diagram for ASR pipeline

| | |
|---|---|
| WER (ngram=2) | 30.41 |
| WER (ngram=4) | 23.01 |

# Experiments - Scenario 1

- Utterance of three speakers of different native languages (Hindi, Telugu, Malayalam)
- Captured across 4 sentences (total of 12 unique sentences).
- **Parameters:** ngram=2

| | speaker_name | accent | actual_text | transcribed_text | wd_in_actual_text | wd_in_corpus | wd_correctly_uttered |
|---|---|---|---|---|---|---|---|
| 0 | siddharth | hindi_eng | we have been discussing about newton's laws of motion | we have been discussing about new audience they'll ask cost margin | 9 | 8 | 5 |
| 1 | siddharth | hindi_eng | if you want a rainbow you gotta put up with the rain | i guess you've want to funding will do what are adopted the same | 12 | 11 | 2 |
| 2 | siddharth | hindi_eng | eagles do not take flight lessons from chickens | biggest the markets like the essence some statements | 8 | 6 | 0 |
| 3 | siddharth | hindi_eng | i will do my homework on time everyday | it's high single platform will contain every day | 8 | 8 | 1 |
| 4 | rohit | telugu_eng | conversational ai has large scope for research | and litigation and being fast last call persistence | 7 | 6 | 0 |
| 5 | rohit | telugu_eng | data science engineers always have a decent income | the data science begin is always have a decent income | 8 | 8 | 7 |
| 6 | rohit | telugu_eng | mistakes are always forgivable if one has the courage to admit them | a mistake stock on based on you know that is one has that primate selected | 12 | 11 | 2 |
| 7 | rohit | telugu_eng | i will do my homework on time everyday | i to michael will content every day | 8 | 8 | 2 |
| 8 | deepu | malayalam_eng | nlp stands for natural language processing | and increased transform actually manage costs | 6 | 5 | 0 |
| 9 | deepu | malayalam_eng | trust no one be the only one | i just low oil be that we've won | 7 | 7 | 1 |
| 10 | deepu | malayalam_eng | summer school has been great experience | the summer support faster main page 6 patients | 6 | 6 | 1 |
| 11 | deepu | malayalam_eng | i will do my homework on time everyday | private label will perform activity | 8 | 8 | 1 |

deepu1.wav

rohit2.wav

siddharth3.wav

# Experiments - Scenario 2

- Utterance of three speakers of different native languages (Hindi, Telugu, Malayalam)
- Captured across 3 unique sentences.
- **Parameters:** ngram=2
- **Sentences:** "we will study computer science",
- "math is an important subject",
- "I will discuss the key topics for exam over the next few hours"

| | speaker_name | accent | actual_text | transcribed_text | word_count_actual | word_in_corpus | wd_correctly_uttered |
|---|---|---|---|---|---|---|---|
| 0 | siddharth | hindi_eng | we will study computer science | i mean is steady on u s banks | 5 | 5 | 0 |
| 1 | siddharth | hindi_eng | math is an important subject | i know that this is an important subject | 5 | 5 | 4 |
| 2 | siddharth | hindi_eng | i will discuss the key topics for exam over th... | i believe will discuss our team topics so i th... | 13 | 13 | 8 |
| 3 | rohit | telugu_eng | we will study computer science | we remain steady compugen since | 5 | 5 | 1 |
| 4 | rohit | telugu_eng | math is an important subject | that east and important subject | 5 | 5 | 2 |
| 5 | rohit | telugu_eng | i will discuss the key topics for exam over th... | i mean then discuss the key topics plot again ... | 13 | 13 | 8 |
| 6 | deepu | malayalam_eng | we will study computer science | we remain very steady contour of science | 5 | 5 | 2 |
| 7 | deepu | malayalam_eng | math is an important subject | and throughout the east and is often such as | 5 | 5 | 1 |
| 8 | deepu | malayalam_eng | i will discuss the key topics for exam over th... | high to discuss key topics forex some forward ... | 13 | 13 | 5 |

deepu2.wav

rohit2.wav

siddharth2.wav

# Experiments - Result Comparison

| accent | wd_in_actual_text | wd_in_corpus | wd_correctly_uttered |
|---|---|---|---|
| hindi_eng | 37 | 33 | 8 |
| malayalam_eng | 27 | 26 | 3 |
| telugu_eng | 35 | 33 | 11 |

Speaker accent-wise accuracy distribution for Scenario 1

| accent | word_count_actual | word_in_corpus | wd_correctly_uttered |
|---|---|---|---|
| hindi_eng | 23 | 23 | 12 |
| malayalam_eng | 23 | 23 | 8 |
| telugu_eng | 23 | 23 | 11 |

Speaker accent-wise accuracy distribution for Scenario 2

# Future Work

- There continues to be scope for further improvement in accuracy and Word Error Rate.
- In the future, we hope to apply various combinations of architectures and parameters.
- For the current model different possible parameters in form of n-gram values can be tested.
- Besides other architectures such as Gaussian Mixed Models (GMMs) and Time Delay Neural Networks (T-DNNs) may be attempted for the Indian accented English case.
- Work with different accent lines.

**Setup the project in your system, convert your voice to text** github.com/sidgupta234/Indian_English_ASR

# References

Babu, L. B., George, A., Sreelakshmi, K. R., & Mary, L. (2018). Continuous Speech Recognition System for Malayalam Language Using Kaldi. 2018 International Conference on Emerging Trends and Innovations In Engineering And Technological Research (ICETIETR), 1–4. https://doi.org/10.1109/ICETIETR.2018.8529045

Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Park, J., Kyubyong &. Kim. (2019). G2pE. In GitHub repository. GitHub. https://github.com/Kyubyong/g2p

Povey, D., Ghoshal, A., Boulianne, G., Goel, N., Hannemann, M., Qian, Y., Schwarz, P., & Stemmer, G. (2011). The kaldi speech recognition toolkit. In IEEE 2011 Workshop.

Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. INTERSPEECH.

# THANK YOU
# /ˈθæŋk ju/
# TH AE NG K Y UW