Building an ASR system for Indic Languages - Indian English

Siddharth Gupta

Indian Institute of Technology Delhi, New Delhi

Rohit Reddy

Amrita Vishwa Vidyapeetham, Coimbatore

Deepu C

ICFOSS, Thiruvananthapuram

IIITH Advanced NLP Summer School 2022

Mirishkar S Ganesh (Mentor)

International Institute of Information Technology, Hyderabad

July 8, 2022

Contents

Abstract

With the leaps in computational power and ever-increasing amount of structured speech transcribed data to avail, the accuracy of Automated Speech Recognition (ASR) systems has seen substantial improvements over the last few years. Given a large amount of transcribed data, the systems have proven to be capable of performing especially well when speech is produced by native speakers. In cases when a language, say, English, is spoken by L2 speakers, there may be a heavy influence of their native language on their accent when they speak English; the scenario can make it difficult for an ASR system to make correct transcriptions. The accent influence for building an ASR system is a big challenge for speakers of a country like India- one of the most linguistically diverse countries, which has a large number of multilingual non-native English speakers. 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.

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

Keywords: Automated Speech Recognition, Hidden Markov Models, Kaldi, Indian English

About the Dataset

The dataset used for training the HMM model consists of 39341 .wav 16Khz mono-channeled audio files (each file being considered as having one utterance) from NPTEL lecture videos delivered in English. The shortest utterance is of 4 words duration, while the longest one is 67 words long. The mean utterance length is 24.18 words with a standard deviation equaling 8.520 (more statistics can be consulted from Table 1). The distribution is further visually represented as a histogram (Figure 1). Part of the Speech analysis was carried out using SpaCy's(Honnibal & Montani, 2017) "en_core_web_sm" model (see Figure 2). Word-frequency table for the dataset corpus can be read in Figure 3.

Approach

Popular speech processing tool Kaldi (Povey et al., 2011) - with the help of its official documentation was used to create the ASR model for continuous Indian English speech transcription. Figure 4(Babu et al., 2018) gives an overview of the modeling process. Here we briefly explain the steps used in the process.

- Text preprocessing: The text corpus was preprocessed through two prominent steps: lowercasing and substituting punctuations (except apostrophes) with a space.
- Lexicon generation: Generated a lexicon for every word in the text corpus. This step was achieved through g2p library(Park, 2019). For eg. the phoneme for ‘thank you’ would be ‘TH AE NG K Y UW’. Read Table 2 for a portion of the lexicon file.
- Train/Test split: An 80:20 split was performed on the utterance dataset for train and test split.
- Audio feature extraction: MFCCs (Mel-Frequency Cepstral Coefficients) are extracted from .wav audio files and CMVN (Cepstral mean and variance normalization) is applied to perform normalization.
- Training: HMM modeling is performed on extracted and normalized MFCCs.
- Language Model: Creating n-gram (n=1, 2, 4) using SRI Language Modelling Toolkit(Stolcke, 2002).
- Decoding: Calculates the likelihood of words forming a sequence.
- Identifying utterance: After decoding, with the help language model and HMM model utterance is identified.

Experiments

1. 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, test cases are mentioned in Table 3a. The detailed outcome of the experiment can be observed in table 3b.

2. Scenario 2: Utterance of three speakers of different native languages (Hindi, Telugu, Malayalam) captured across 3 unique sentences. Parameters: ngram=2, test cases are mentioned in Table 3b. The detailed outcome of the experiment can be observed in table 4b.

Results

1. Scenario 1: Speaker accuracy (number of words present in both actual audio and generated transcription) over different sentences has been clubbed and produced in table 3c. It can be inferred that the current model is performing best for the Telugu accented speaker, followed by the Hindi and Malayalam speakers viz.
2. Scenario 2: Speaker accuracy over the same three sentences (comparative study) has been clubbed and produced in table 4c. It can be inferred that the current model is performing best for the Hindi accented speaker, followed by Telugu and Malayalam speakers viz.

Future Scope

On further analysis of the current state of our ASR model, we have inferred that 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.

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*.

Tables

Table 1

Average number of words in a sentence

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

Table 2

Lexicon File

affect AE F EH K T

affected AH F EH K T AH D

affecting AH F EH K T IH NG

affects AE F EH K T S

affiliate AH F IH L IY EY T

affiliated AH F IH L IY EY T AH D

affiliates AH F IH L IY AH T S

affiliation AH F IH L IY EY SH AH N

affinity AH F IH N AH T IY

affirm AH F ER M

affirmation AE F ER M EY SH AH N

affirmed AH F ER M D

affirming AH F ER M IH NG

affirms AH F ER M Z

affluent AE F L UW AH N T

afford AH F AO R D

affordability AH F AO R D AH B IH L AH T IY

affordable AH F AO R D AH B AH L

affordably AH F AO R D AH B L IY

afforded AH F AO R D AH D

affording AH F AO R D IH NG

Table 3a

Testcases for Scenario 1

‘we have been discussing about newton's laws of motion ‘,

'if you want a rainbow you gotta put up with the rain',

'eagles do not take flight lessons from chickens',

'i will do my homework on time everyday',

'conversational ai has large scope for research',

'data science engineers always have a decent income',

'mistakes are always forgivable if one has the courage to admit them',

'i will do my homework on time everyday',

'nlp stands for natural language processing',

'trust no one be the only one',

'summer school has been great experience',

'i will do my homework on time everyday'

Table 3b

Accuracy description table for Scenario 1

	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

Table 3c

Speaker accent-wise accuracy distribution for Scenario 1

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

Table 4a

Testcases for Scenario 2

"we will study computer science",

"math is an important subject",

"I will discuss the key topics for exam over the next few hours",

"We will study computer science",

"Math is an important subject",

"I will discuss the key topics for exam over the next few hours",

"We will study computer science",

"Math is an important subject",

"I will discuss the key topics for exam over the next few hours"

Table 4b

Accuracy description table for Scenario 2

	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

Table 4c

Speaker accent-wise accuracy distribution for Scenario 2

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

Figures

Figure 1

Frequency distribution of words in sentence

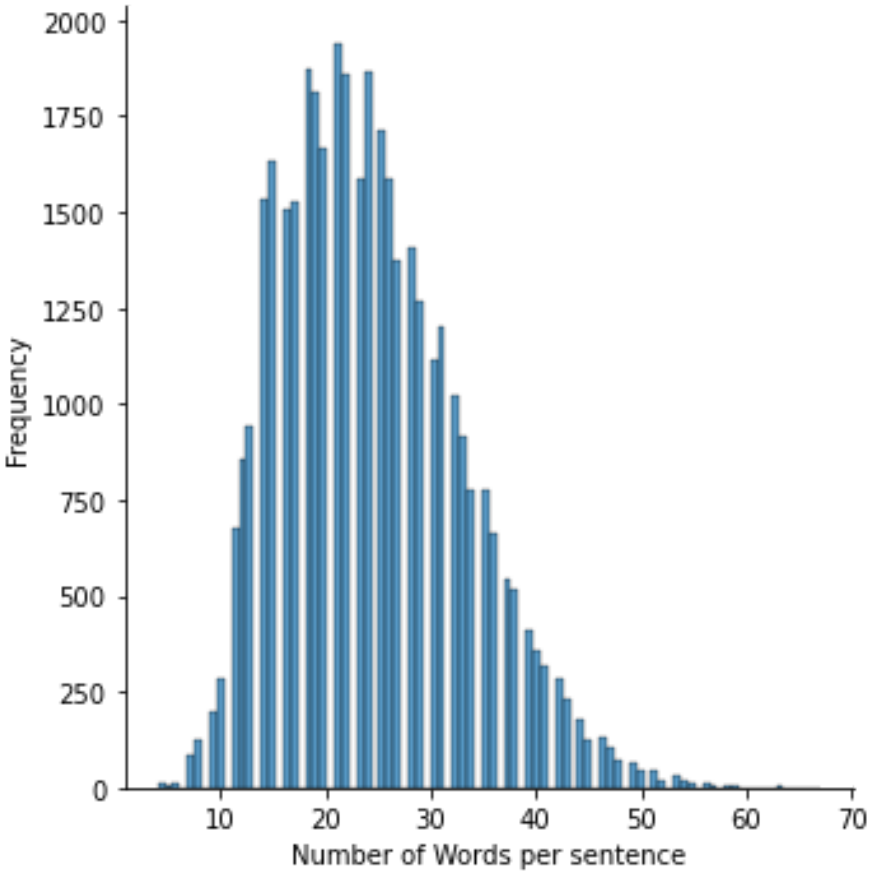

Figure 2

Frequency distribution of Part of Speech

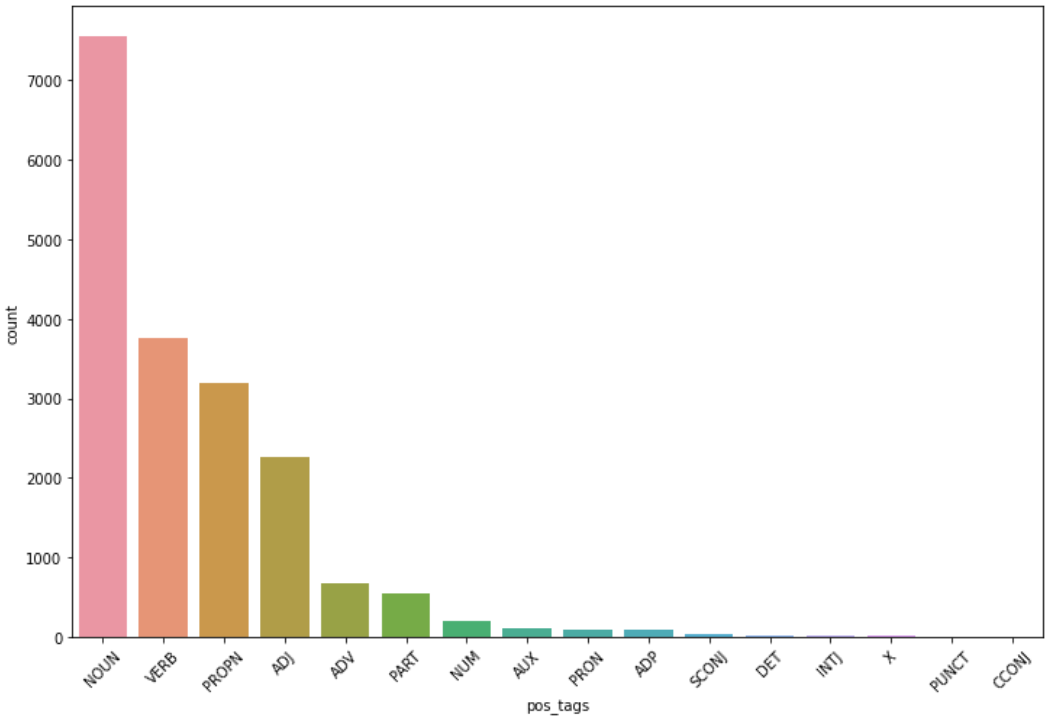

Figure 3

Frequency distribution of vocabulary in corpus

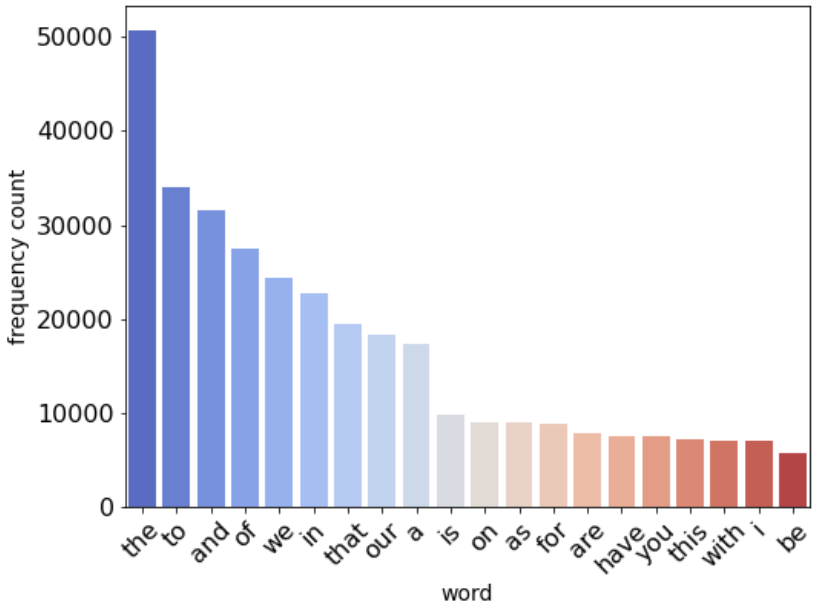

Figure 4

Block diagram for ASR modeling