Automatic Speech Recognition

Group 1

Pravalika Kommagouni
Vijay Bhasker Kodam
Varsha Loya Ravathkar
Varma MRKS
Mahojjwal

Group 2

Gagan Kawde
Anil Garikipati
Nav Pallav
Harika Kasimahanthi
Divya Kasimahanthi

Contents

Speech and Voice Recognition Market to Rise Exponentially at a CAGR of 19.8%

Project Overview

The project involves developing an Automatic speech recognition model for Telugu/Hindi languages.

Automatic speech recognition is the capability of a computer system to decipher spoken words and phrases from audio and transcribe them into written text.

We choose the Kaldi toolkit for training the ASR using some hours of transcribed data.

We are going to adopt HMM (Hidden Markov Model), GMM (Gaussian Mixture Model) and Deep learning models to achieve the Speech-To-Text functionality.

Problem Statement

Industry requires ASR for a broad range of commercial products. Hence, ASR as user interface has become ever more useful and pervasive.

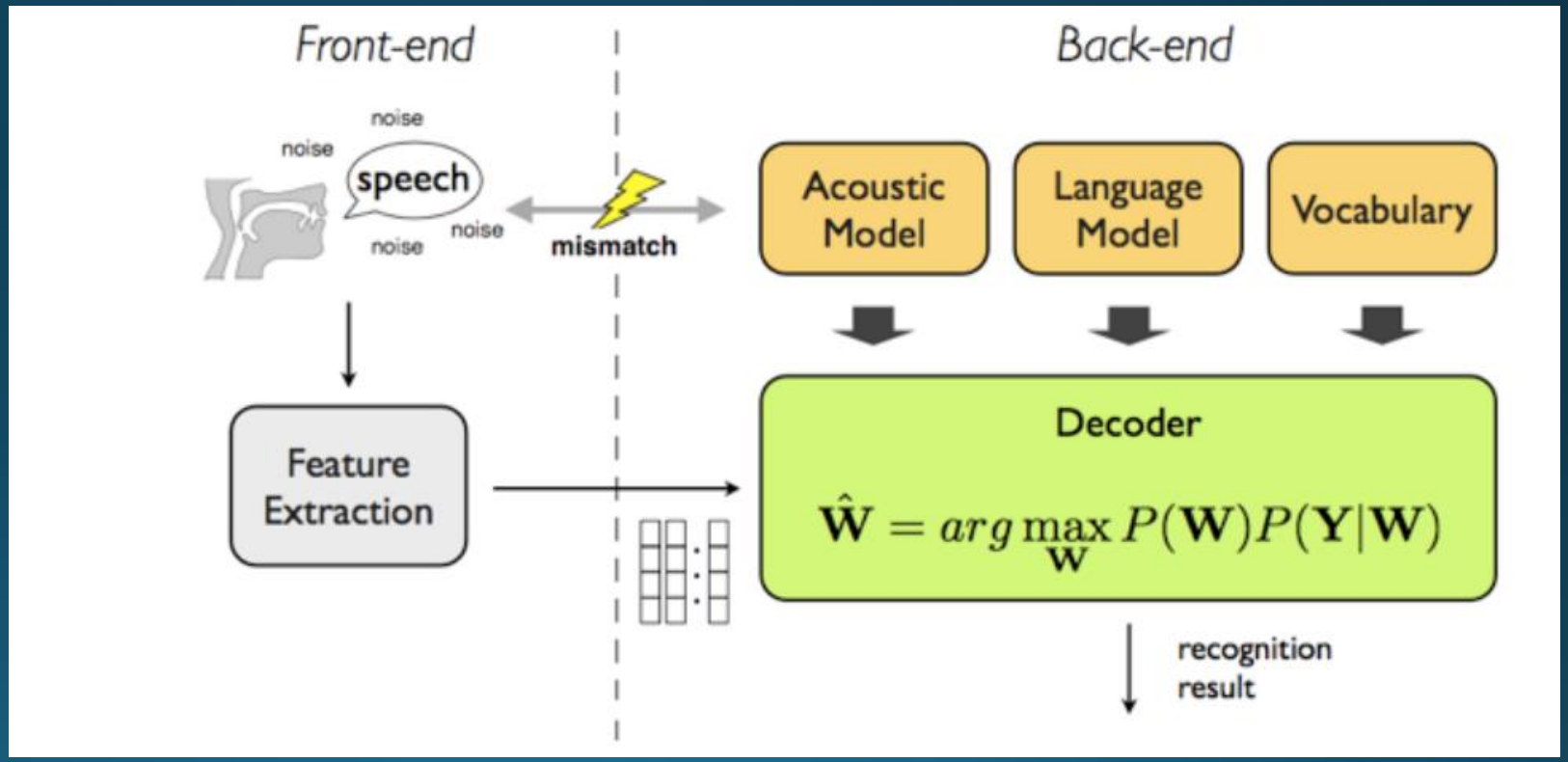
Users can control devices with the help of speech and voice as the audio and text received by such devices automatically gets converted into a machine-friendly format, making it easier for humans to operate devices without having to indulge time and effort by operating other devices such as a mouse, keyboard, and others.

However, reliably recognizing spoken words in realistic acoustic environments is still a challenge.

Our Project focusses on a deployable model for Speech-text Conversion for Native Indian Languages – Telugu and Hindi.

Mathematical Formulation

- X denotes the feature vector derived from acoustic data and W denotes a string of words; therefore, the recognizer is capable of picking up the most preferred string.
- P(W|X) denotes the probability that W are uttered given that X is detected.
- The argmax function is referred to the most likely input that leads to maximum output.



Database Statistics

Group 1

Language Chosen : Telugu

Train Data Size :53 hours

Test Data Size : 6 hours

Dev Data Size : 1 hour

Facts : Telugu is the 4th most spoken language in India.

Group 2

Language Chosen : Hindi

Train Data Size : ~29 Hours

Test Data Size : ~10 hours

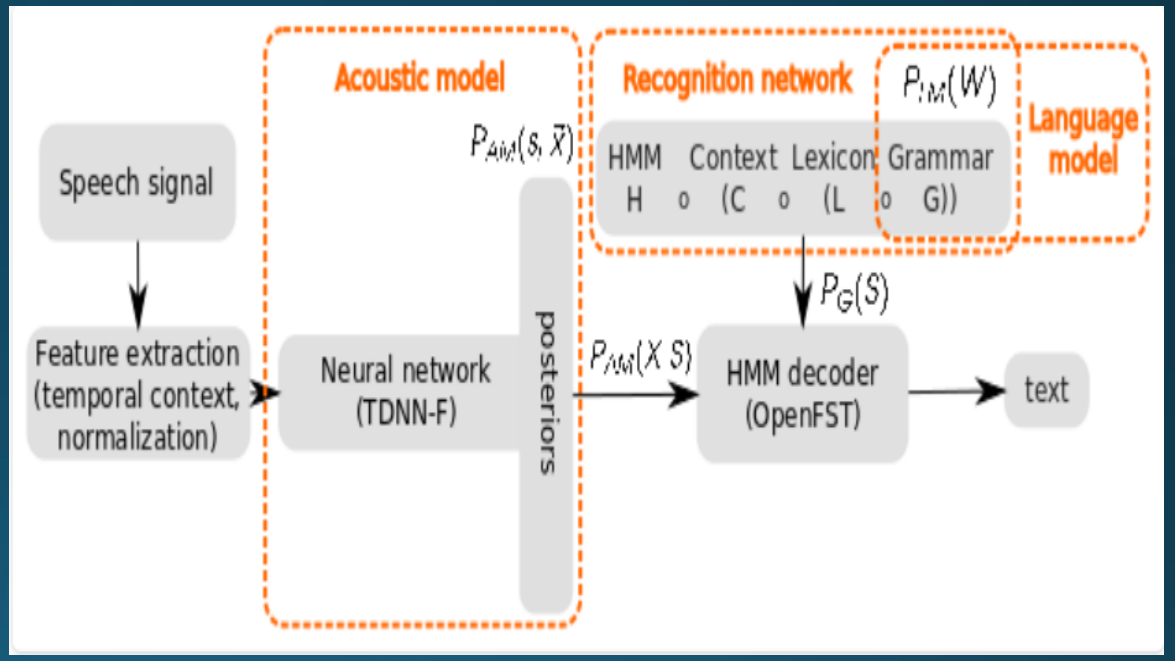
Dev Data Size : ~3 hours

Facts : Hindi is the most spoken language in India.

Methodologies

The central component is the HMM decoder, which explores the search space defined by the Recognition network H ◦ C ◦ L ◦ G. Internally, the partial paths have costs computed from acoustic model scores Pam(X|S) and graph scores PG(S), where S is the hypothesized state-sequence of the partial path from the beginning of the utterance.

In the decoding formula (W) we see that we search for such state sequence S that has the maximal score, and we read the corresponding word sequence W' by using the operator wrds(.). For optimal Word Error Rate (WER) performance, two hyper-parameters need to be identified, the acoustic scale κ and language-model scale ρ.



$$\hat{W} = \text{wrds}\left(\underset{S}{\text{argmax}}\, P_{AM}(\mathbf{X}|S)^{\kappa}\, P_G(S)^{\rho}\right)$$

Stages & Timelines

Data Collection/Preparation	Model Build/Test	Deployment
• Collect the Audio Files • Pre-processing (Audio/Text)	• Feature Extraction • Feature Normalization • Build/Train the Model • Test the Model with the Test Data • Blind Test	• REST API • WebSocket(TBD)
20th Nov – 27th Nov,2022	**27th Nov,2022 – 7th Jan,2023**	**7th Jan – 11th Feb,2023**

References

- https://www.atco2.org/news/automatic-speech-recognition-how-it-works --> slide with methodologies
- https://www.prnewswire.com/in/news-releases/speech-and-voice-recognition-market-to-rise-exponentially-at-a-cagr-of-19-8-cloud-based-segment-to-emerge-dominant-says-fortune-business-insights-828795577.html
- https://www.esat.kuleuven.be/psi/spraak/theses/08-09-en/MDT.php
- https://www.atco2.org/news/automatic-speech-recognition-how-it-works