Speech Understanding

Assignment 1

Speech-to-Text Conversion

Akansha Gautam (M23CSA506)

Anchit Mulye (M23CSA507)

Introduction

- Speech-to-Text Conversion, also called Automatic Speech Recognition (ASR)
- It is the process of processing the speech signals and converting the spoken language into text format using computational models

- Provides real-team captions for videos and meetings, allowing people with hearing impairments to access spoken content
- Enables people to understand the audio content better if there's some foreign accent present in the audio
- Heavily used by customer care agents in automating call transcriptions
- Speech-to-text conversion with machine translation facilitates people with different geographies to communicate better

Strengths

Whisper:

- Multilingual Support
- Transcribe + Translate
- Open Source
- No language specific tuning

Nova:

- Faster
- Speaker diarization, filter words support
- Better in pre recorded and real time transcription

Limitations

Whisper:

- No Real Time Transcription
- Huge Size
- Accents and Dialects Issue
- Extreme Background Noise

Nova:

- Only English language Support
- Not available on open source

Results

Model	WER*
Whisper	13.61%
Nova	15.48%

- Nova has WER of 15.48% on the **LJ Speech Dataset**, available on Kaggle
- Whisper has WER of 13.61% on the same dataset
- Whisper is trained on huge multilingual audio data of 680,000 hrs and due to this it performs better

* Both the models were ran locally, code is available in the github repo.

Problems

- Most of the models are trained in high-resource languages like English and struggle with regional languages.
- Difficult for the STT models to extract the linguistic context present in the audio signal in noisy environments.
- Models struggle because of different accents and pitch variations present in audio signal.

Opportunities

- Models can be trained on multilingual datasets
- We can optimize the current models by processing the audio file in smaller chunks parallelly, it'll help with the latency
- STT models lack context awareness, leading to misinterpretation of homophones. Combining speech + video (lip reading) + text for improved accuracy.

Thank You