SPEECH TO TEXT PROJECT REPORT

Joy Olusanya

August 26, 2024

1 Introduction

Speech to text (STT) or speech recognition enables human computers to interact by translating spoken language into written text. This technology has evolved significantly, leveraging artificial intelligence (AI) advancements, particularly neural networks and deep learning. Speech to Text models are capable of handling diverse speech patterns, accents, dialects, and background noise, making them more accurate and reliable[3].

2 Model Selection

2.1 Background

Whisper model was selected due to its state of art performance in speech to text task [2]. Whisper is a model that build based on transformer encoder-decoder architecture and is trained in weakly supervised on 680000 hours of multilingual and multitasking[4].

2.2 Why Whisper model?

For this project, Whisper model was chosen because of its accuracy in handling diverse English accents and also multilingual data.

3 Audio Data Selection

The audio data used for this project was sourced from Librivox recordings public domain. The audio file was in mp3 format. Whisper model was used for transcribing the pre-processed audio data into text.

4 Model Transcription Performance

The Word Error Rate (WER) is a metric for evaluating the performance of Automatic Speech Recognition (ASR) systems. WER is most common metric[1], It measures the percentage of words that are incorrectly recognized by the ASR model. The Speech to Text model transcription text was compared with the Ground Truth (The manually transcribe text). The Word Error Rate result output was 0.0093 which indicates that the model performs extremely well in accurately transcribing speech to text.

References

[1] Foteini Filippidou and Lefteris Moussiades. A benchmarking of ibm, google and wit automatic speech recognition systems. In *Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings, Part I 16*, pages 73–82. Springer, 2020.

[2] Riefkyanov Surya Adia Pratama and Agit Amrullah. Analysis of whisper automatic speech recognition performance on low resource language. *Jurnal Pilar Nusa Mandiri*, 20(1):1–8, 2024.

[3] Ayushi Trivedi, Navya Pant, Pinal Shah, Simran Sonik, and Supriya Agrawal. Speech to text and text to speech recognition systems-areview. *IOSR J. Comput. Eng*, 20(2):36–43, 2018.

[4] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.