Milestone II - Training and Model Selection for Project (HackArizona)

Kanit Mann (kanitmann@arizona.edu)

Tanishk Singh (tanishksingh@arizona.edu)

University of Arizona

INFO 511: Introduction to Data Science

Prof. Angela Cruze

February 25th, 2025

Milestone II

Building on our initial groundwork from Milestone 1, our focus for Milestone 2 is on the practical application of the Common Voice (Common Voice) and CoVoST(GitHub - facebookresearch/covost: CoVoST: A Large-Scale Multilingual Speech-To-Text Translation Corpus (CC0 Licensed)) datasets to develop a peer-to-peer translation tool. This tool aims to facilitate real-time cross-language translation, aligning with our broader goal of delivering impactful solutions through data management and visualization.

Data Identification and Acquisition

The Common Voice and CoVoST datasets were selected for their extensive multilingual speech data, which is crucial for training robust end-to-end speech-to-text (ST) models, which will in turn act as bare metal frame for speech-to-speech conversion.

Methods and Initial Processing and Handling Missing Data

We have initiated exploratory data analysis (EDA) to understand the dataset's structure, focusing on language distribution and speaker diversity. This analysis helps identify potential biases and ensures the data's representativeness.

Data processing includes cleaning and normalizing audio files and transcriptions, a critical step for maintaining data quality. We are employing noise reduction techniques and audio normalization to address inconsistencies in audio quality. Additionally, we are aligning audio data with translations to streamline the integration into our ST model training pipeline.

We encountered challenges such as missing transcriptions, poor quality transcriptions and variable audio quality. These issues are being addressed through data augmentation techniques and cross-referencing with supplementary datasets. Our approach ensures that the data remains comprehensive and reliable for model training.

Use Cases

The primary use case for this data is the development of a real-time peer-to-peer translation tool which can be used for seamless communication between two people without need of a common language.

In summary, our work on Milestone 2 builds on the foundation laid in Milestone 1, focusing on the practical application of the Common Voice and CoVoST datasets to develop a real-time translation tool. Our structured approach to data handling and project management ensures that we are well positioned to deliver a solution that meets our objectives.