

Week 2: Data Collection & Initial Preprocessing

TASK 1: Collect & Upload Dataset Files (Kyle & Wissam)

Goal: Gather and upload speech dataset samples so we can start processing audio features.

What Needs to Be Done:

- 1. Download dataset files from Common Voice and L2-ARCTIC.**
 - **Common Voice:** Extract English speech samples (preferably from American, British, and non-native Vietnamese speakers).
 - **L2-ARCTIC:** Extract Vietnamese-accented English speech samples.
- 2. Organize dataset files based on accent type.**

Create three subfolders inside data/ in GitHub

data/

|— american/

|— british/

|— vietnamese/

- 3. Ensure all files meet these requirements:**
 - **Format:** .wav
 - **Sample rate:** 16kHz
 - Mono-channel audio (if stereo, convert to mono)
 - **Length:** 2-10 seconds per sample
 - **If conversion is needed, use SoX:** `sox input.wav -r 16000 -c 1 output.wav`
- 4. Upload dataset files to GitHub (data/ folder).**
 - At least 10-20 files per accent must be uploaded first so the next tasks can begin.

TASK 2: Confirm Google Colab Setup (James)

Goal: Set up a shared Google Colab notebook for testing and running feature extraction.

What Needs to Be Done:

1. **Install required libraries in Colab (if not already installed):** !pip install librosa numpy pandas scikit-learn matplotlib tensorflow torch torchaudio
2. **Create a Colab Notebook with:**
 - A test script to load an audio file from GitHub.
 - A command to print basic audio properties (duration, sample rate).
 - A waveform plot of the loaded audio file.
3. **Upload the Colab notebook to GitHub (notebooks/ folder).**
4. **Share the Google Colab link with the team so everyone can access and test it.**

TASK 3: Start Feature Extraction Testing (Tai)

Goal: Extract and visualize key speech features (MFCCs, Pitch, Formants) to understand accent differences.

What Needs to Be Done:

- 1. Load dataset samples in Google Colab (once James shares the link).**
- 2. Extract speech features using librosa:**
 - Compute MFCCs (Mel-Frequency Cepstral Coefficients).
 - Extract Pitch and Formants to analyze how accents differ.
 - Plot features visualizations for different accents.
- 3. Save extracted features as .csv or .npy files for training.**
- 4. Upload the feature extraction script to GitHub (scripts/ folder).**