Assignment 1: Forced Alignment using Montreal Forced Aligner (MFA)

Objective

Set up and execute a complete forced alignment pipeline using the Montreal Forced Aligner (MFA) tool. And to understand how automatic alignment works between speech audio and phonetic transcription.

What is Forced Alignment?

Forced alignment is the process of automatically matching an audio recording with its corresponding text transcription at the word and phoneme level.

It determines when each word or sound begins and ends in the speech signal.

In simple terms, if you know what was said (the transcript), forced alignment helps you find when it was said in the audio.

Suppose we have the following:

Audio:

A speaker says — "Hello world"

Transcript:

HELLO WORLD

Pronunciation dictionary:

HELLO HH AH L OW

WORLD W ER L D

A forced aligner (like MFA) analyzes the audio and produces a TextGrid file, which marks word and phoneme boundaries (start time, and end time), for example:

Words:

0.00 – 0.45 HELLO

0.45 – 0.90 WORLD

Phones:

0.00 – 0.10 HH

0.10 – 0.25 AH

0.25 – 0.40 L

0.40 – 0.45 OW

0.45 – 0.55 W

0.55 – 0.70 ER

0.70 – 0.85 L

0.85 – 0.90 D

Dataset:

A pre-selected dataset containing audio files and corresponding text transcripts will be provided. Each transcript corresponds to the spoken content in the audio file (typically one utterance per file). You are required to use only the provided dataset for all experiments.

Task Overview:

You are required to:

1. Set up the MFA environment
2. Install Montreal Forced Aligner on your system
3. Prepare the data
    a. Organize the dataset into the MFA-required format
4. Select or train a pronunciation dictionary
    a. Use an existing MFA dictionary (e.g., english_us_arpa) or
    b. Train your own dictionary from transcripts if required (using a G2P model).
5. Run forced alignment
6. Inspect and analyze the output
    a. Check generated TextGrid files using Praat software.
    b. Identify how phoneme and word boundarie**s** are aligned.
    c. Observe any errors or mismatches in alignment (e.g., skipped phonemes, timing offsets).
7. How to handle Out of vocabulary (OOV) words? And implement the solution.

Submission Instructions:

· GitHub Repository: Push all scripts, setup instructions, and outputs to a public GitHub repository.

· README: Include clear steps to install MFA, prepare the dataset, and run the alignment (with example commands).

· Outputs: TextGrid files.

· Report: Summarizing model/dictionary used, sample alignment visualization, and key observations in before and after OOV handling.

· Access: Ensure all links (GitHub and Drive) are publicly accessible for evaluation.

Resources of data:

You can find the audio files in ‘wav’ folder and corresponding transcripts in the 'transcripts' folder.