Assignment Report

Forced Alignment using Montreal Forced Aligner (MFA)

Abstract:

This report presents a forced alignment experiment using Montreal Forced Aligner (MFA) on English speech data. Audio-transcript pairs were processed to generate TextGrid files with word- and phoneme-level boundaries. The alignments were inspected in Praat, showing high accuracy with minor deviations in rapid speech. Out-of-vocabulary words were handled using dictionary extension and G2P models. Results confirm MFA's effectiveness in automatic speech-text alignment.

Objective:

The objective of this assignment is to perform forced alignment between speech audio files and their corresponding text transcripts using the Montreal Forced Aligner (MFA). The task also aims to understand how word-level and phoneme-level boundaries are automatically aligned with the audio signal and to analyze the alignment output using Praat.

What is Forced Alignment:

Forced alignment is the process of automatically aligning a given speech recording with its known transcription. It determines the start and end time of each word and phoneme in the audio.

Dataset Description:

The dataset consists of speech audio files in WAV format and corresponding transcript files in TXT format. Each audio file has a matching transcript file with the same filename.

Tools and Models Used:

Montreal Forced Aligner (MFA), Pretrained English Acoustic Model (MSK / english_mfa), and Praat.

Methodology:

The dataset was prepared in the required MFA format. Forced alignment was performed using a pretrained English acoustic model, which generated TextGrid files containing word- and phoneme-level boundaries.

Implementation:

The forced alignment pipeline was implemented using the Montreal Forced Aligner (MFA) on a local Windows system. MFA version 2.x was used along with a pretrained English acoustic model (english_mfa, also referred to as MSK).

The speech dataset was organized according to MFA requirements, with each audio file (.wav) paired with a corresponding transcript file (.txt) having the same filename. The alignment process generated TextGrid files containing word- and phoneme-level timing information.

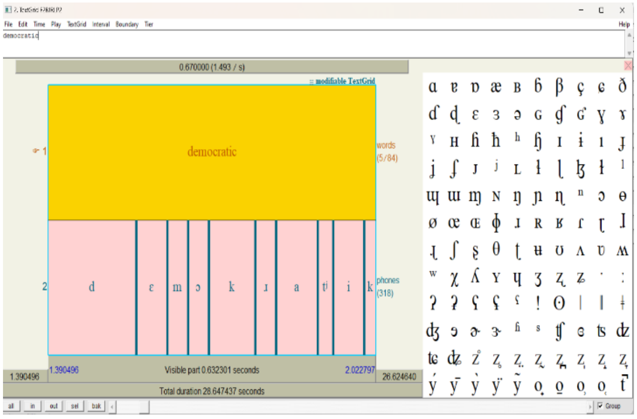
The following command was used to perform forced alignment:

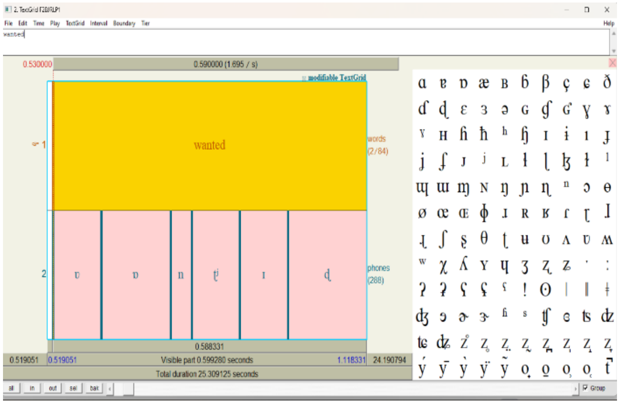
```
mfa align corpus english_mfa english_mfa output
```

After alignment, the generated TextGrid files were opened in Praat along with the corresponding audio files for visualization and analysis.

Inspection and Analysis using Praat:

The generated TextGrid files were opened in Praat along with their corresponding audio files. Word and phoneme boundaries were inspected using waveform and spectrogram views. The alignments were mostly accurate, with minor deviations in fast speech segments.





Out-of-Vocabulary (OOV) Words Handling:

OOV words were handled by extending the pronunciation dictionary or by using a G2P model to generate pronunciations for unseen words, followed by re-running the alignment.

Results:

MFA successfully generated TextGrid files. Visual inspection confirmed accurate alignment between audio and transcripts.

Conclusion:

This assignment demonstrated a complete forced alignment pipeline using MFA and Praat, providing insight into automatic speech-text alignment.

Forced Alignment of Speech Using MFA
by Sameeksha.