1. Introduction

Forced alignment is the process of automatically aligning audio with its corresponding text transcript at the word and phoneme levels. Montréal Forced Aligner (MFA) performs this by using an acoustic model and a pronunciation dictionary. The output is a Praat-compatible TextGrid file containing word and phone boundaries.

This assignment involves installing MFA, preparing the dataset, performing alignment, and analyzing the generated TextGrid files.

2. Dataset Description

The provided dataset consists of:

- A **wav** folder containing speech audio files
- A **transcripts** folder containing corresponding text transcripts
  Each audio file represents a single utterance spoken by one speaker.

The dataset was reorganized into the standard MFA corpus structure prior to alignment.

3. Tools Used

- **Montreal Forced Aligner:** Version 3.3.x
- **Acoustic Model:** `english_us_arpa`
- **Pronunciation Dictionary:** `english_us_arpa`
- **Environment:** Miniconda-based virtual environment
- **Visualization Tool:** Praat (for TextGrid inspection)

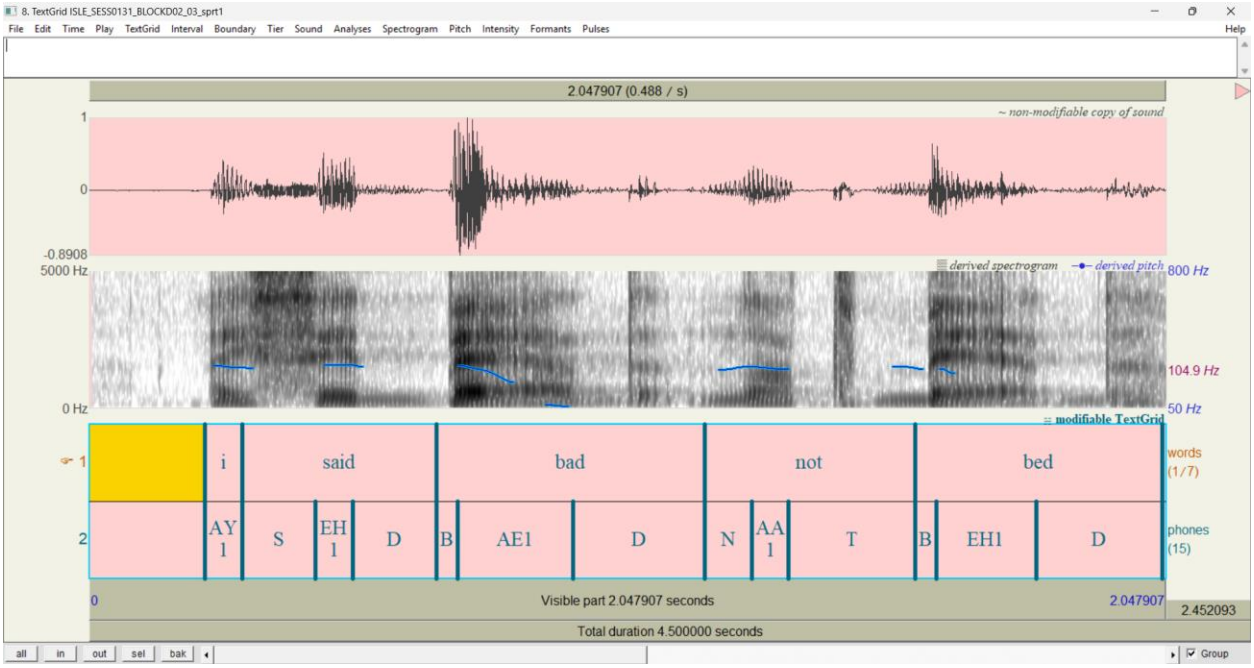
4. Forced Alignment Procedure

The alignment was performed using the following steps:

1. Created and activated a conda environment
2. Installed MFA using pip
3. Organized files into the required MFA corpus format
4. Used the command:
5. `mfa align <corpus_path> english_us_arpa english_us_arpa <output_path>`
6. Generated TextGrid files were opened in Praat for analysis.

5. Alignment Output and TextGrid Analysis

A sample alignment is shown below



TextGrid Analysis:

The TextGrid output contains two key tiers: a **word tier** and a **phoneme (phone) tier**. The word tier correctly segments the utterance *"I said bad not bed"* into individual word intervals, with clear start and end boundaries. The phoneme tier further breaks each word into its corresponding phones based on the MFA dictionary (e.g., *I → AY1*, *said → S EH1 D*, *bad → B AE1 D*, etc.).

Overall, the alignment is mostly accurate. The phoneme boundaries match visible acoustic cues in the waveform and spectrogram—vowel regions appear longer and more intense, while stop consonants (B, D, T) align with short periods of low energy. The transition between *bad* and *not* is well aligned, showing distinct boundaries in both tiers.

However, minor timing shifts are observable. For example, the boundary between **EH1** and **D** in *bed* may start slightly earlier than the visible change in the waveform. Such small mismatches are common due to speaker pronunciation variations, coarticulation, and model assumptions in MFA. Another slight deviation occurs at the onset of *said*, where the **S** boundary overlaps with noise-like frication but does not perfectly match the spectrogram.

These alignment imperfections are expected and reflect natural variability in speech as well as limitations of the generic acoustic model. Despite this, the overall alignment is consistent, interpretable, and suitable for further analysis.

.

6. Observations and Issues

- Alignment is accurate for most words.
- Slight mismatches in boundaries occur due to:
    - speaker variability
    - reduced pronunciation
    - coarticulation effects
    - limitations of the generic acoustic model
- No major errors were found (e.g., missing segments or swapped phones).

7. Conclusion

The Montreal Forced Aligner successfully aligned the provided dataset with high accuracy. The generated TextGrid files offer detailed phonetic segmentation suitable for further linguistic or speech processing tasks. Small alignment errors are normal and stem from speech variability and model assumptions.