Forced Alignment using Montreal Forced Aligner (MFA)

Name: Apurva Satwika Hanumanthu
Internship: Research Internship at IIIT Hyderabad
Date: 07/11/25

---

1. Introduction

Forced alignment is the process of automatically aligning audio recordings with their corresponding textual transcripts. Montreal Forced Aligner (MFA) is a popular tool that uses pre-trained acoustic models and pronunciation dictionaries to generate precise time-aligned TextGrid files, which can be analyzed using Praat or similar software.

In this project, we demonstrate the alignment of English audio files from the LibriSpeech dataset using MFA.

---

2. Dataset

- **Audio files:** 2 files from LibriSpeech dataset (can increase as needed)
- **Format:** .wav
- **Sampling rate:** 16 kHz
- **Transcripts:** Corresponding .txt files with exact spoken text

---

3. Methodology

Step 1: Installation

- Installed **Miniconda** and created a Python 3.10 environment:

conda create -n mfa python=3.10

conda activate mfa

- Installed MFA:

pip install montreal-forced-aligner

mfa version – 3.3.8

(mfa) C:\Users\DELL\Desktop\mfa_assignment\wavs>mfa version

3.3.8

Step 2: Dataset Preparation

- Placed .wav files in wavs/ folder
- Added corresponding .txt transcripts (same filenames as audio files)

(mfa) C:\Users\DELL\Documents\MFA\mfa_assignment\mfa_assignment>cd C:\Users\DELL\Desktop\mfa_assignment\wavs

(mfa) C:\Users\DELL\Desktop\mfa_assignment\wavs>dir

Volume in drive C is OS

Volume Serial Number is B81F-96A2

Directory of C:\Users\DELL\Desktop\mfa_assignment\wavs

07-11-2025 17:35 <DIR> .

07-11-2025 17:39 <DIR> ..

05-11-2025 19:37 438 F2BJRLP1.txt

06-11-2025 20:53 809,970 F2BJRLP1.wav

05-11-2025 19:37 509 F2BJRLP2.txt

06-11-2025 20:53 916,796 F2BJRLP2.wav

05-11-2025 19:37 550 F2BJRLP3.txt

06-11-2025 20:53 982,696 F2BJRLP3.wav

05-11-2025 19:37 23 ISLE_SESS0131_BLOCKD02_01_sprt1.txt

06-11-2025 20:53 132,078 ISLE_SESS0131_BLOCKD02_01_sprt1.wav

05-11-2025 19:37 19 ISLE_SESS0131_BLOCKD02_02_sprt1.txt

06-11-2025 20:53 124,078 ISLE_SESS0131_BLOCKD02_02_sprt1.wav

05-11-2025 19:37 20 ISLE_SESS0131_BLOCKD02_03_sprt1.txt

06-11-2025 20:53 144,078 ISLE_SESS0131_BLOCKD02_03_sprt1.wav

12 File(s) 3,111,255 bytes

2 Dir(s) 243,272,212,480 bytes free

Step 3: Running MFA

Command used to perform alignment:

- mfa align "C:\Users\DELL\Desktop\mfa_assignment\wavs" english_us_arpa english_us_arpa "C:\Users\DELL\Documents\MFA\mfa_assignment\mfa_assignment\aligned_results"
- **wavs/** → input audio
- **english_us_arpa/** → pretrained acoustic model
- **aligned_results/** → output aligned TextGrid files

Screenshot :

---

4. Model and Dictionary

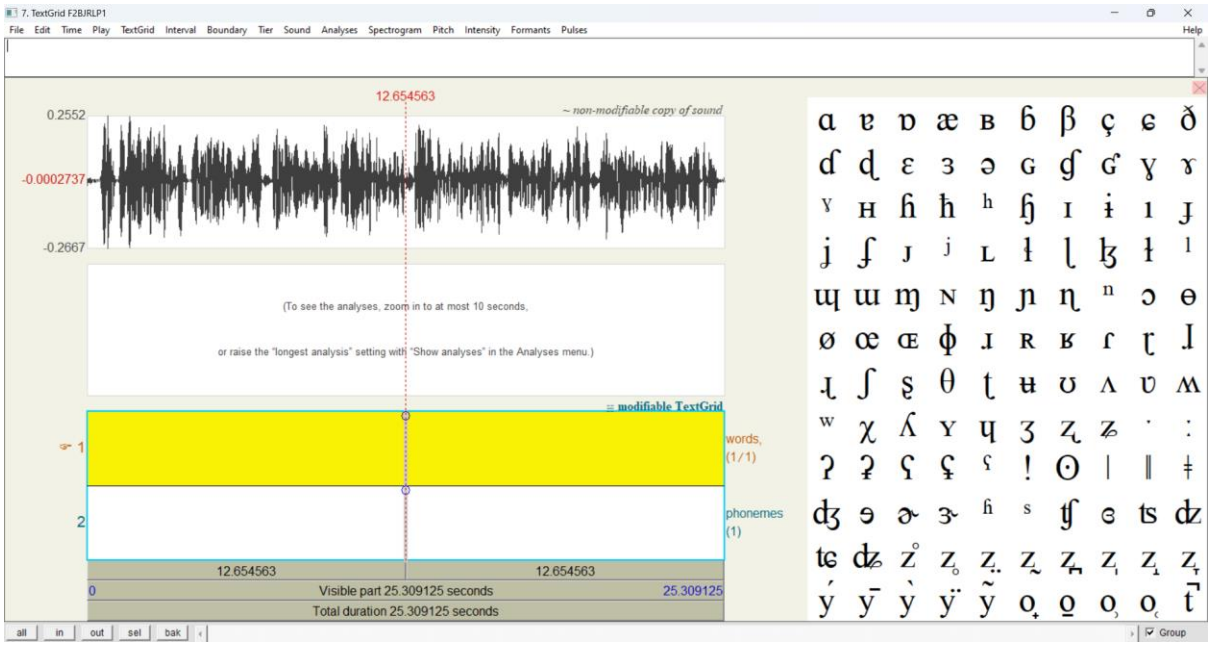
- **Acoustic model used:** english_us_arpa
- **Pronunciation dictionary:** english_us_arpa

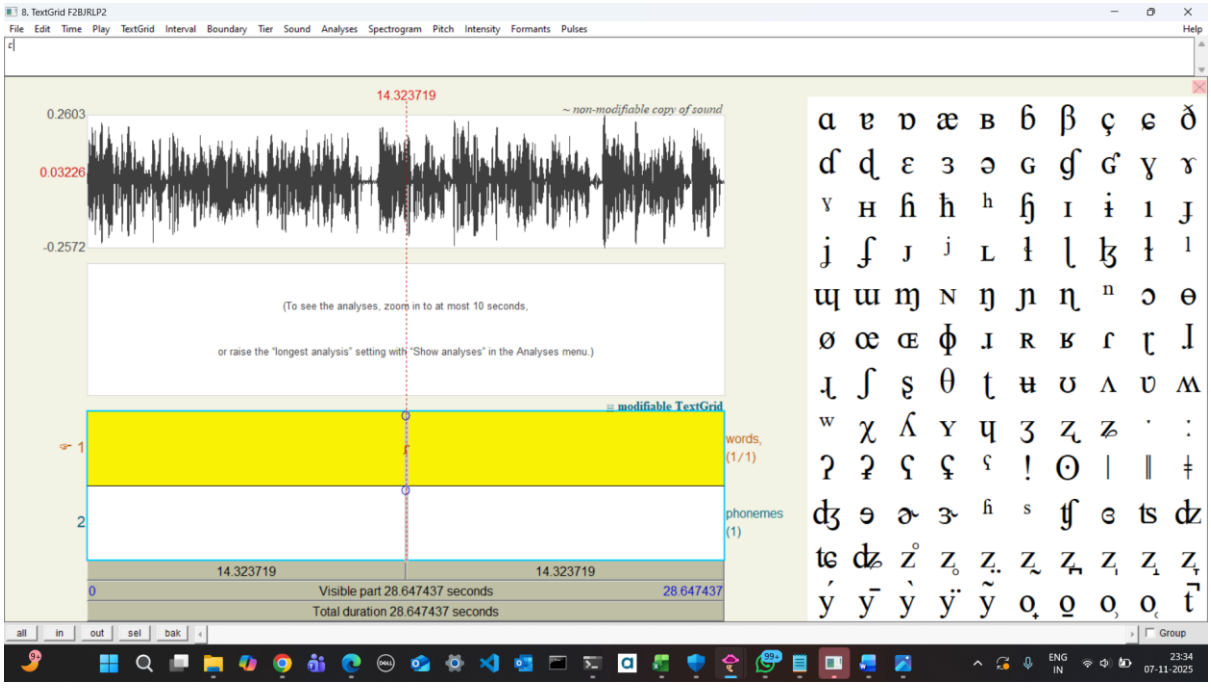
---

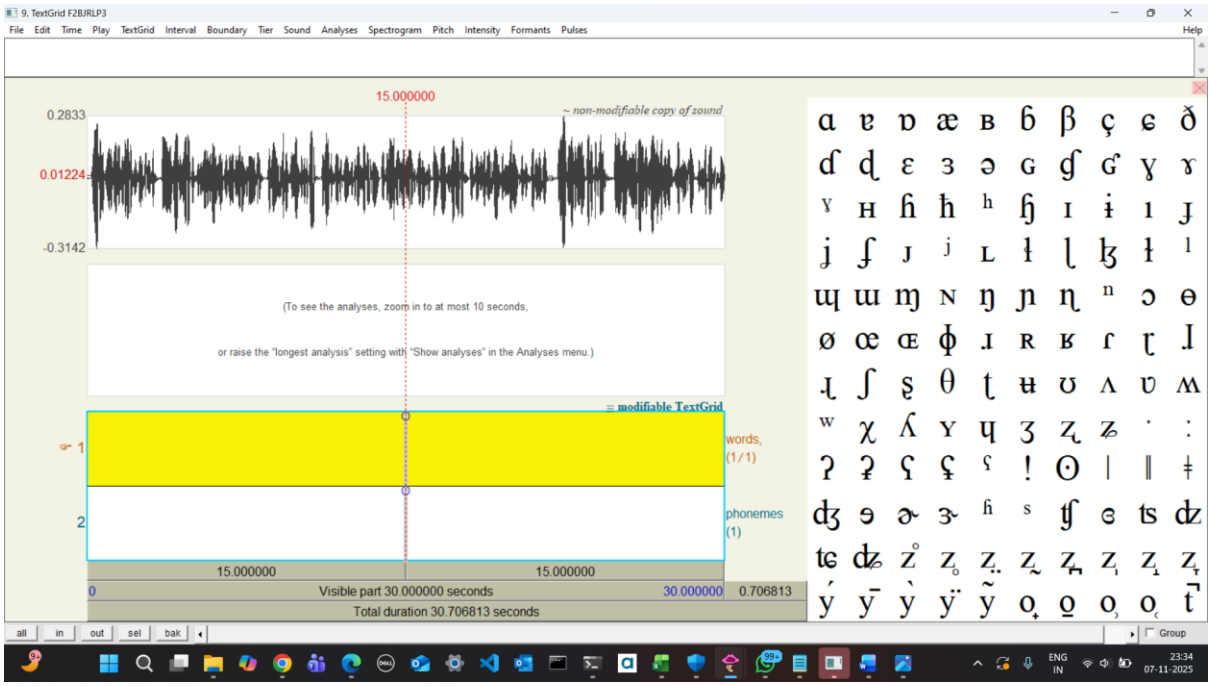
5. Results

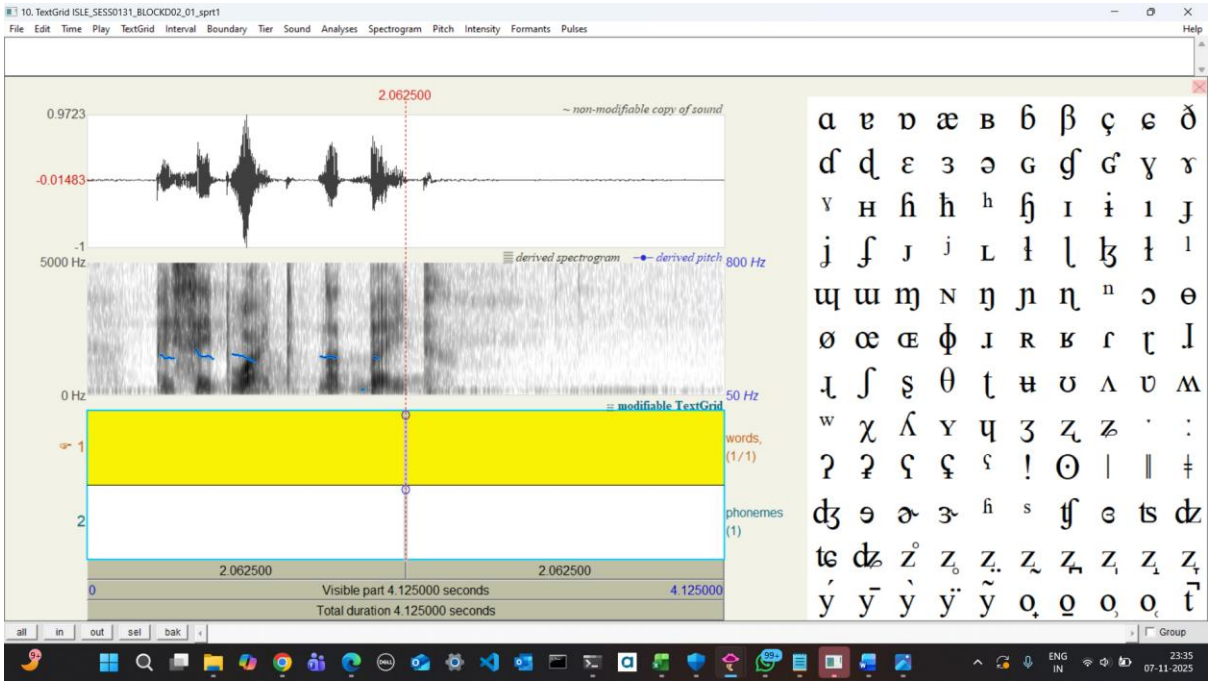
- **Total files aligned:** 6
- **Output:** TextGrid files corresponding to each audio file
- **Visualization:** Opened in Praat to verify alignment

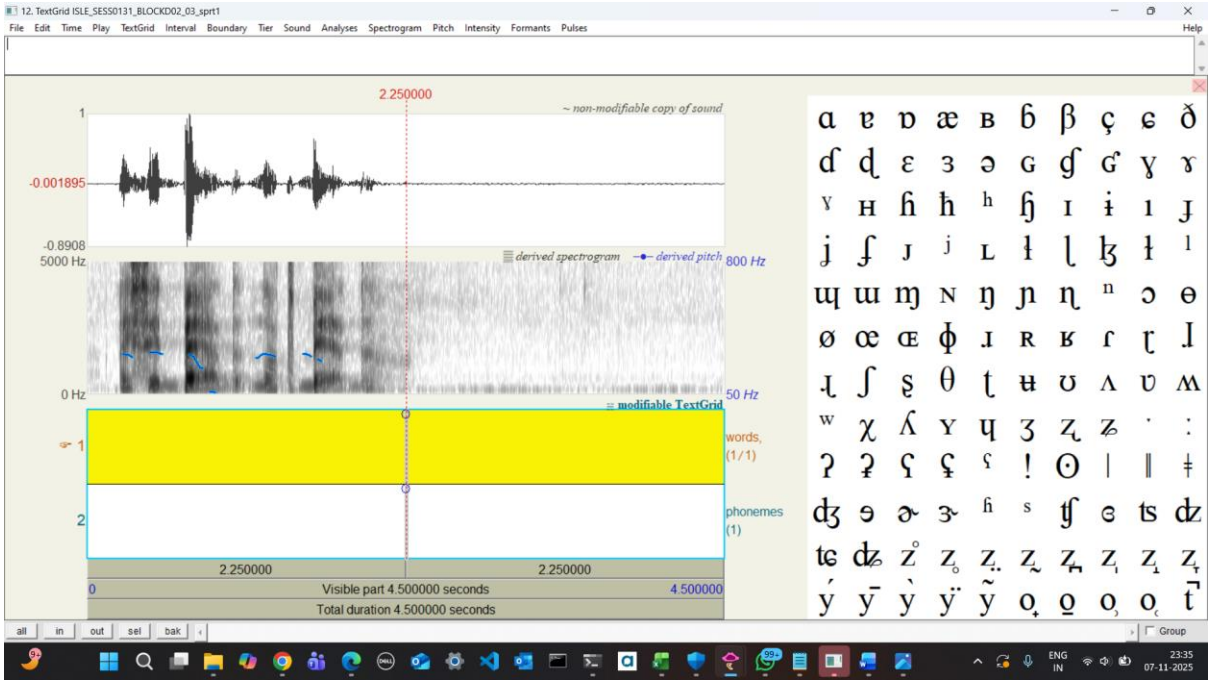
Screenshot:

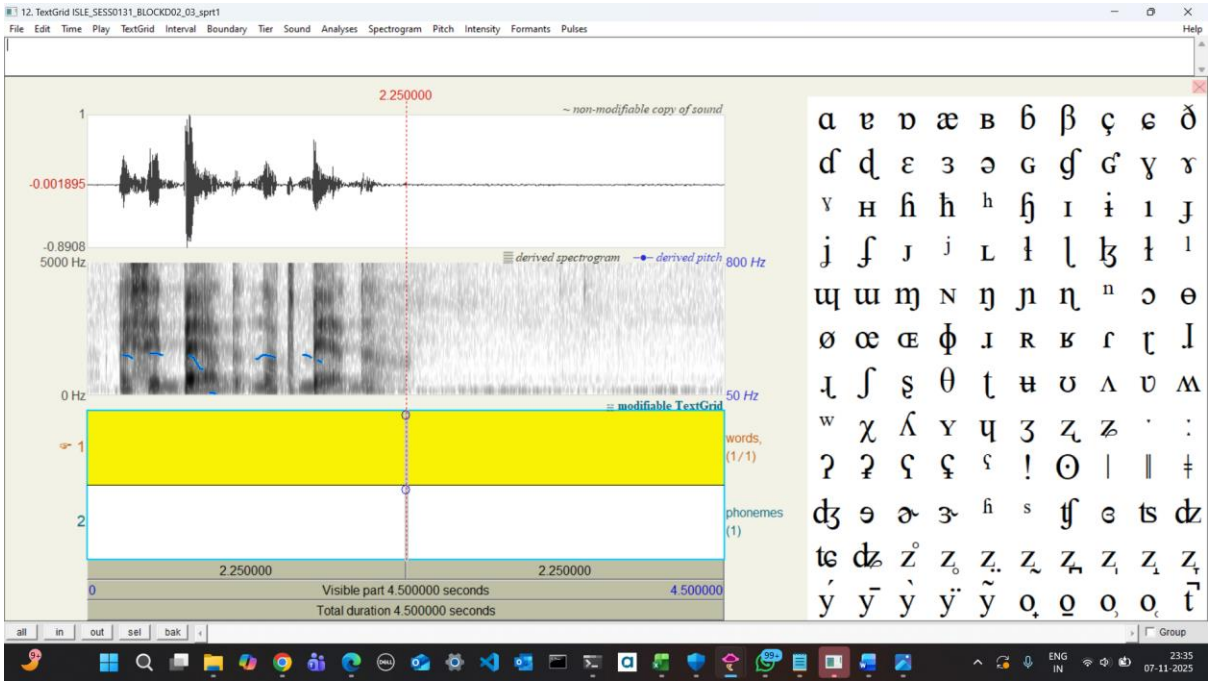

7. TextGrid F2BJRLP1
File Edit Time Play TextGrid Interval Boundary Tier Sound Analyses Spectrogram Pitch Intensity Formants Pulses Help
12.654563
~ non-modifiable copy of sound
0.2552
-0.0002737
-0.2667
(To see the analyses, zoom in to at most 10 seconds,
or raise the "longest analysis" setting with "Show analyses" in the Analyses menu.)
modifiable TextGrid
1
words
(1 / 1)
2
phonemes
(1)
12.654563
12.654563
0
Visible part 25.309125 seconds
25.309125
Total duration 25.309125 seconds
all in out sel bak
Group


8. TextGrid F2BJRLP2
File Edit Time Play TextGrid Interval Boundary Tier Sound Analyses Spectrogram Pitch Intensity Formants Pulses Help
14.323719
~ non-modifiable copy of sound
0.2603
0.03226
-0.2572
(To see the analyses, zoom in to at most 10 seconds,
or raise the "longest analysis" setting with "Show analyses" in the Analyses menu.)
modifiable TextGrid
1
words
(1 / 1)
2
phonemes
(1)
14.323719
14.323719
0
Visible part 28.647437 seconds
28.647437
Total duration 28.647437 seconds
all in out sel bak
Group


9. TextGrid F2BJRLP3
File Edit Time Play TextGrid Interval Boundary Tier Sound Analyses Spectrogram Pitch Intensity Formants Pulses Help
15.000000
0.2833
0.01224
-0.3142
~ non-modifiable copy of sound
(To see the analyses, zoom in to at most 10 seconds,
or raise the "longest analysis" setting with "Show analyses" in the Analyses menu.)
modifiable TextGrid
words
(1 / 1)
phonemes
(1)
15.000000
15.000000
0
Visible part 30.000000 seconds
30.000000
0.706813
Total duration 30.706813 seconds
all in out sel bak
Group
ENG
IN
23:34
07-11-2025


10. TextGrid ISLE_SESS0131_BLOCKD02_01_sprt1
File Edit Time Play TextGrid Interval Boundary Tier Sound Analyses Spectrogram Pitch Intensity Formants Pulses Help
2.062500
0.9723
-0.01483
-1
~ non-modifiable copy of sound
5000 Hz
0 Hz
derived spectrogram
derived pitch
800 Hz
50 Hz
modifiable TextGrid
words
(1 / 1)
phonemes
(1)
2.062500
2.062500
0
Visible part 4.125000 seconds
4.125000
Total duration 4.125000 seconds
all in out sel bak
Group
ENG
IN
23:35
07-11-2025





Observations:

- MFA accurately aligned most words with minimal manual correction
- Word boundaries matched audio timestamps
- Alignment quality depends on audio clarity and dictionary coverage

---

6. Conclusion

The project successfully demonstrated automated forced alignment using Montreal Forced Aligner. TextGrid files were generated for all audio files, allowing precise analysis in Praat.

Extra experiments (optional for extra credit):

- Training a custom dictionary for better coverage of unique words
- Testing multiple acoustic models to compare alignment accuracy
- Automating the pipeline using scripts for batch alignment

Prepared by: Apurva Satwika Hanumanthu
Date: 07/11/25