📄 Research Internship Assignment Report: Forced Alignment using MFA

Applicant Name	Padmasri Ambati
Email Address	padmasriambati2004@gmail.com
GitHub Repository	**https://github.com/PADMASRIAMBATI/Montreal-Forced-Aligner**
Drive Link	**https://drive.google.com/drive/folders/1BGrC0WBOHTCjlkv3bhs_xbq45z2obmjO?usp=sharing**
Assignment	Forced Alignment using Montreal Forced Aligner (MFA)

1. 🎯 Assignment Goal (Objective)

The main goal of this assignment was to build a system for Forced Alignment using the Montreal Forced Aligner (MFA) tool. Forced alignment is the automatic process of precisely matching a spoken audio file with its written text transcription, pinpointing the exact start and end times of every word and phoneme (the individual sounds of speech).

This exercise helps in understanding how computers can segment speech data accurately, which is a key step in large-scale speech-to-speech translation projects.

2. ⚙️ Setup and Data Preparation

A. Environment Setup

MFA was installed on the system using Miniconda to ensure all required dependencies were managed correctly in an isolated environment named mfa.

Installation Commands:

`conda create -n mfa python=3.9`

`conda activate mfa`

`pip install montreal-forced-aligner`

B. Dataset Organization

The necessary audio and transcript files were sourced from the provided Google Drive link. For MFA to work correctly, the files were organized into a single input directory, mfa_assignment_data, following this crucial rule:

- Each audio file (.wav) and its corresponding transcript file (.txt) must share the exact same base name.

File Type	Example File Names	Location in Input Folder
Audio (.wav)	F2BJRLP1.wav, ISLE_SESS0131_...sprt1.wav	/converted_corpus
Transcript (.txt)	F2BJRLP1.txt, ISLE_SESS0131_...sprt1.txt	/converted_corpus

3. 🖥️ Alignment Execution

A. Model and Dictionary Used

To align the English speech, a readily available and highly accurate pre-trained setup within MFA was used:

- **Acoustic Model (AM):** english_us_arpa
- **Pronunciation Dictionary:** english_us_arpa

The dictionary uses the ARPABET standard to represent English sounds.

B. Execution Command

The forced alignment was executed using the following command structure:

`mfa align converted_corpus english_us_arpa english_us_arpa output_alignments`

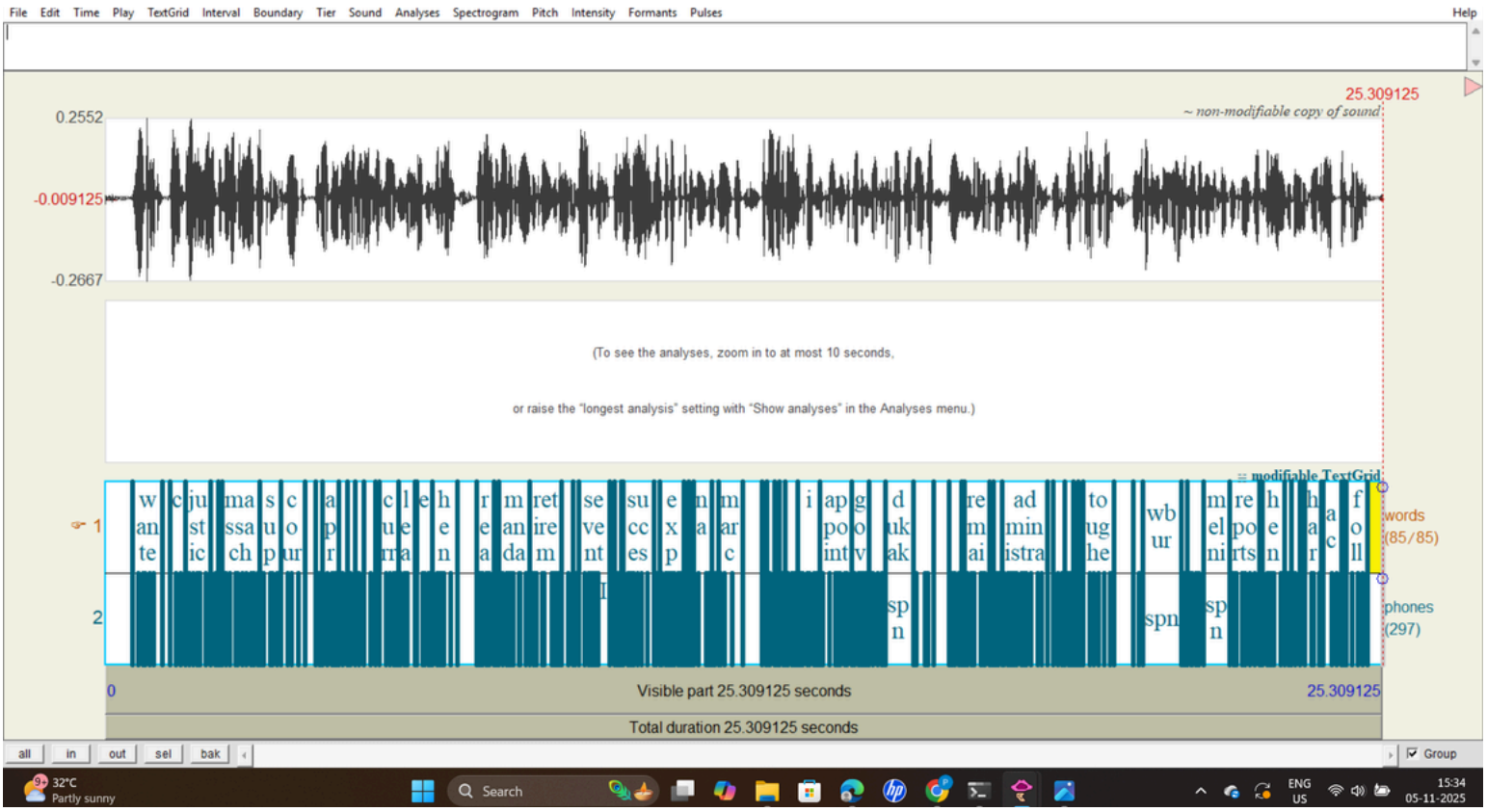
C. Output

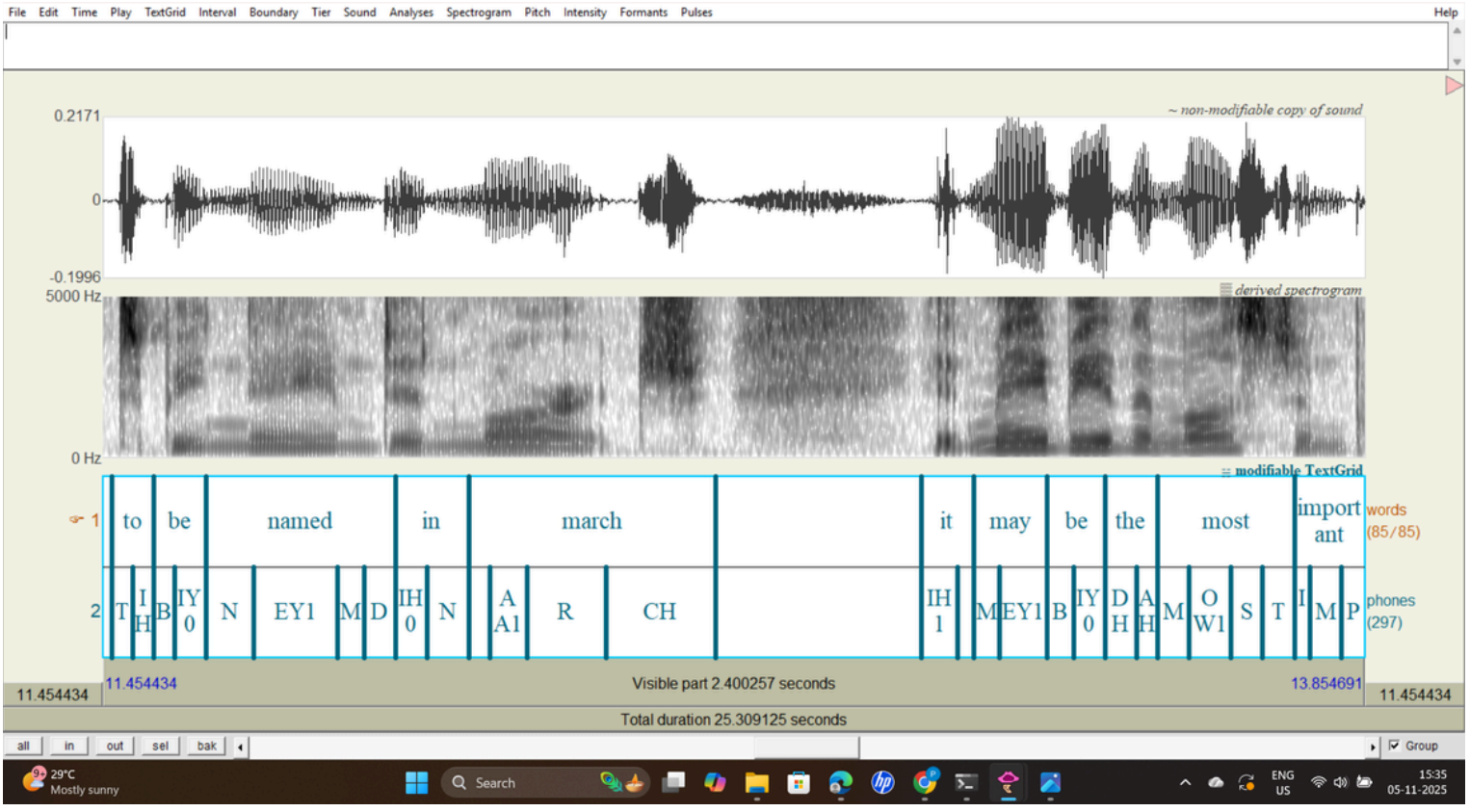
The tool successfully processed all files and generated the time-aligned labels in the form of TextGrid files in the specified output folder.

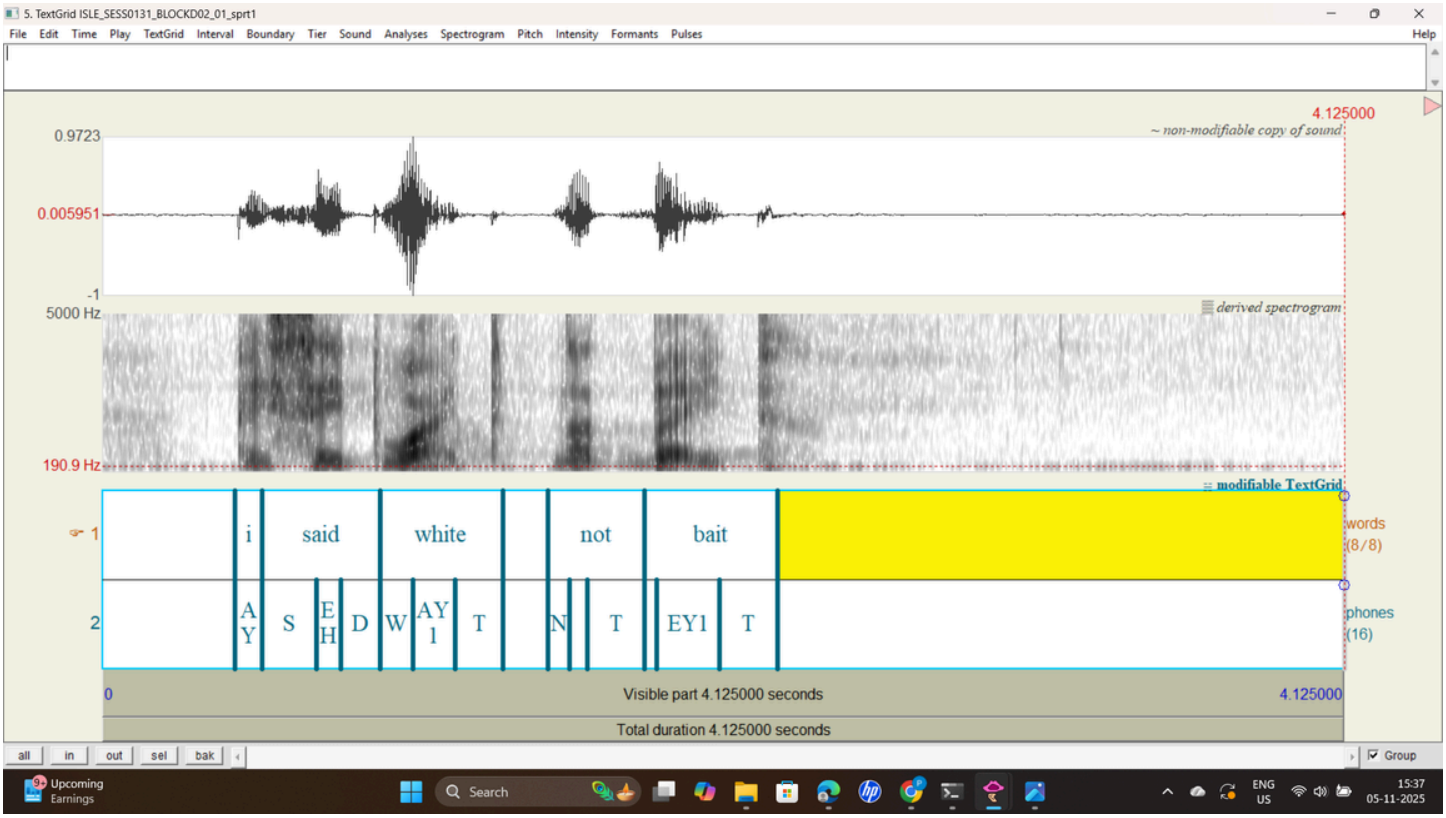
4. 🔬 Analysis and Observations (Praat Inspection)

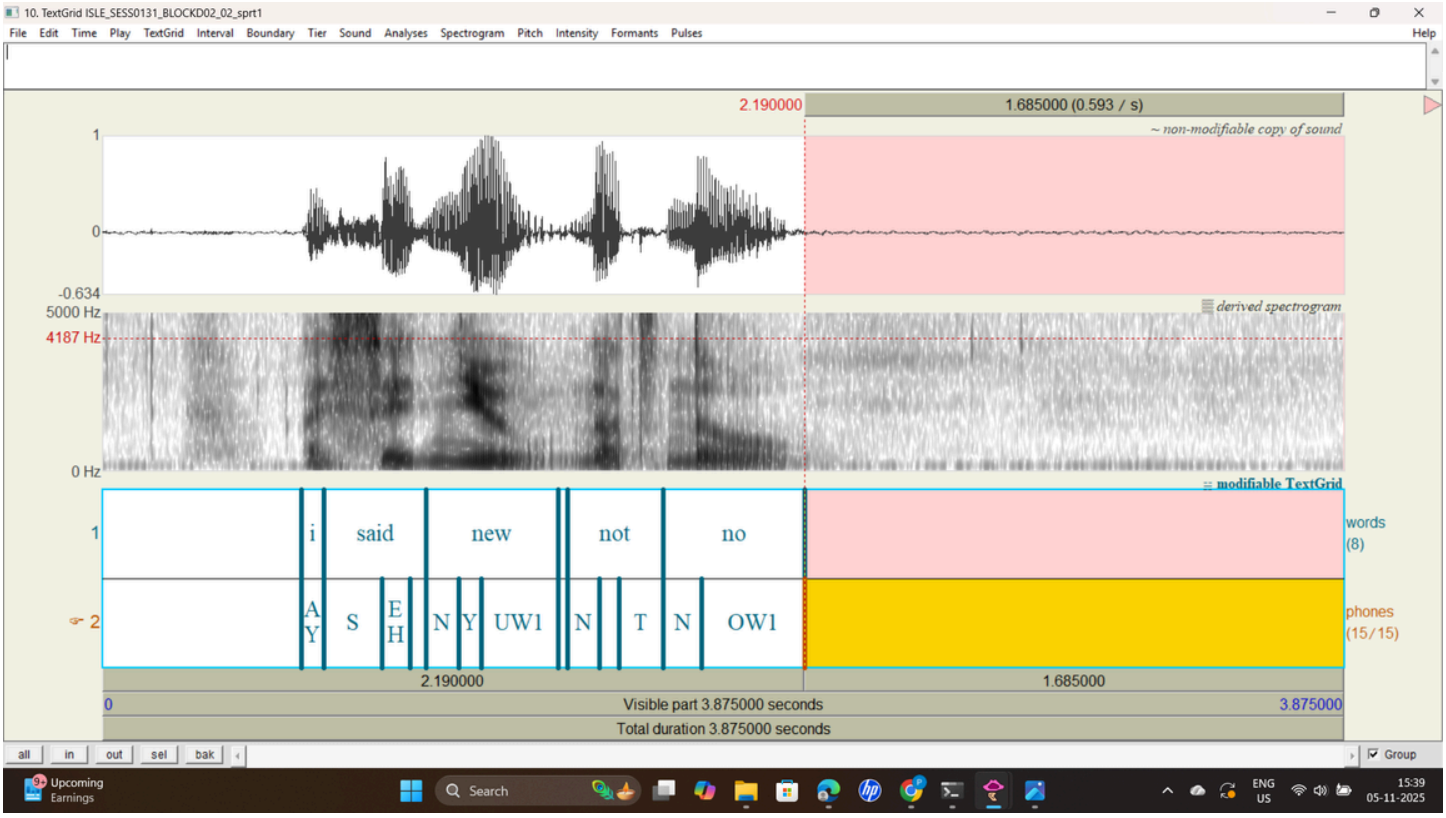
The resulting TextGrid files were opened in the speech analysis software Praat alongside their original audio files to visually verify the alignment accuracy.

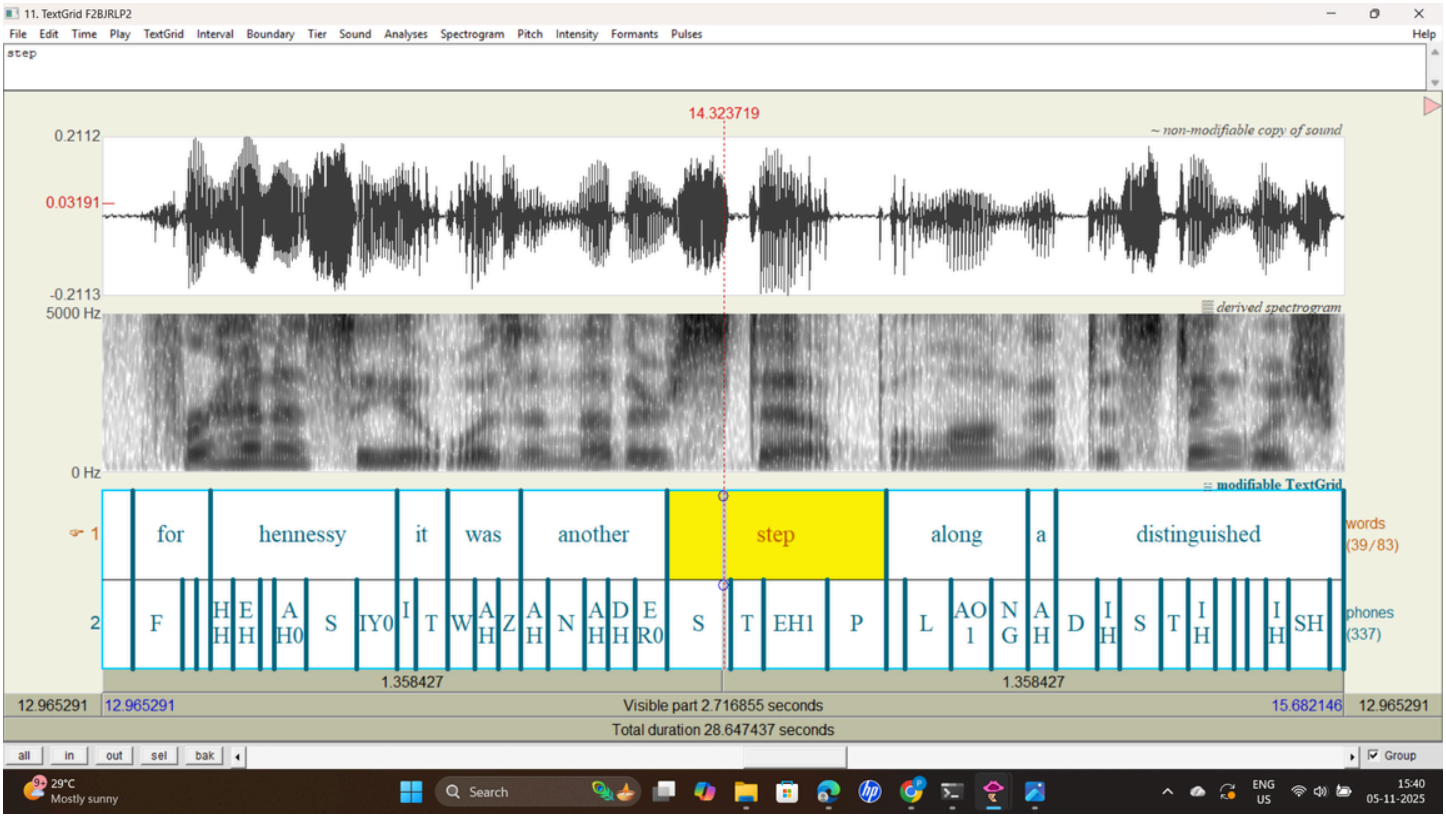
Sample Utterance Analyzed: F2BJRLP1

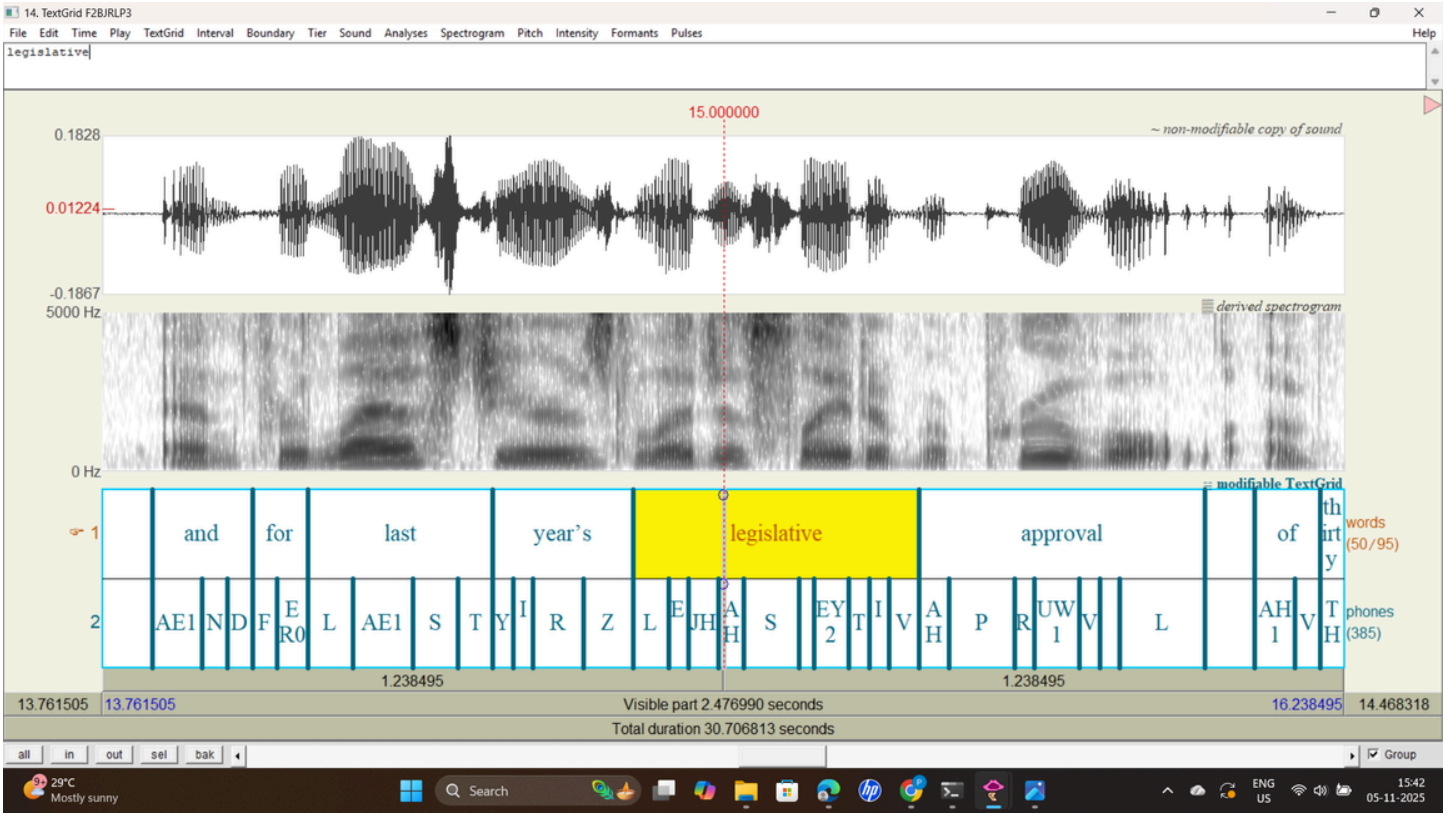

File Edit Time Play TextGrid Interval Boundary Tier Sound Analyses Spectrogram Pitch Intensity Formants Pulses Help
25.309125
~ non-modifiable copy of sound
0.2552
-0.009125
-0.2667
(To see the analyses, zoom in to at most 10 seconds,
or raise the "longest analysis" setting with "Show analyses" in the Analyses menu.)
modifiable TextGrid
words (85/85)
phones (297)
0
Visible part 25.309125 seconds
25.309125
Total duration 25.309125 seconds
all in out sel bak
Group
32°C Partly sunny
Search
ENG US
15:34 05-11-2025

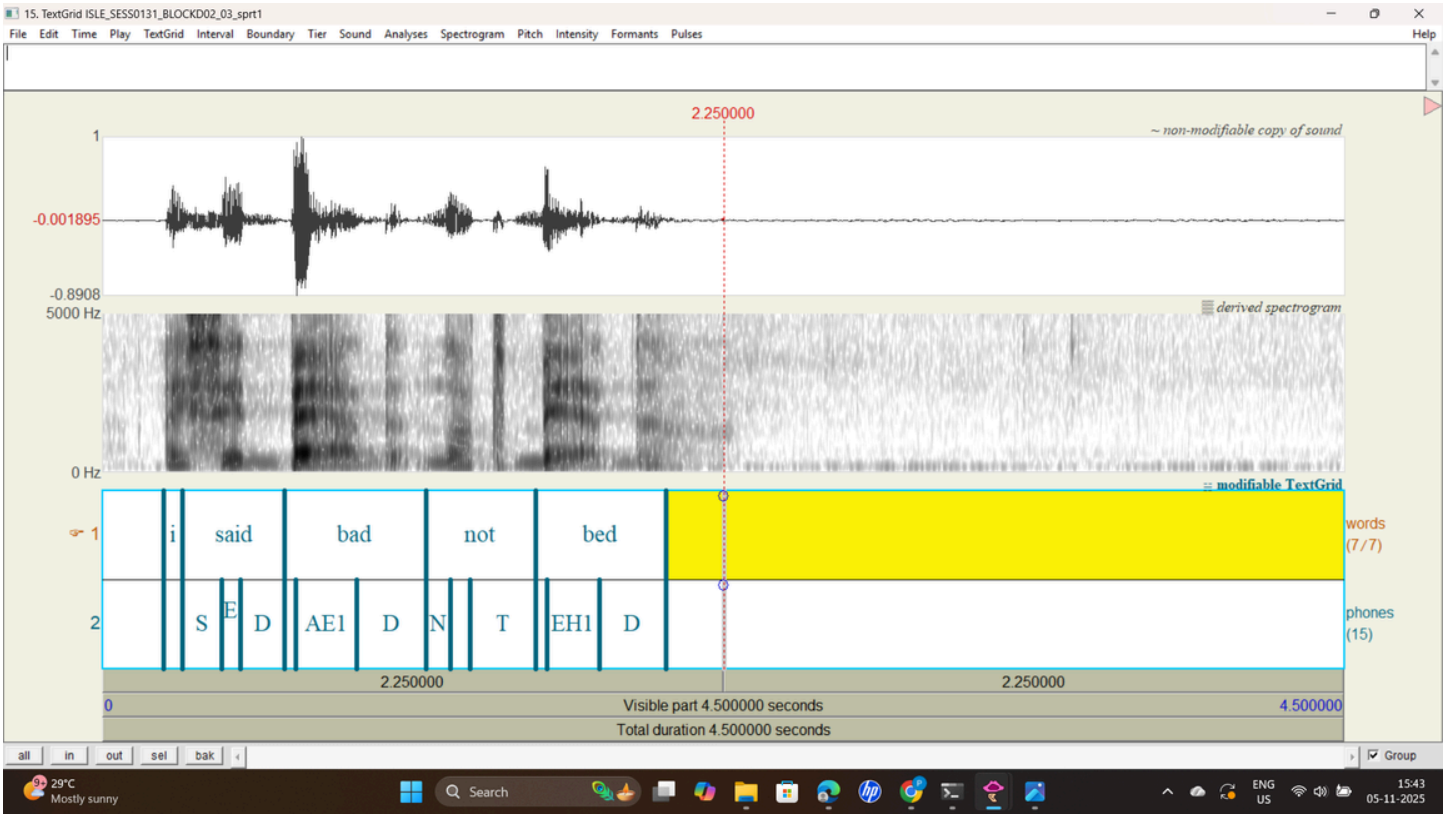

File Edit Time Play TextGrid Interval Boundary Tier Sound Analyses Spectrogram Pitch Intensity Formants Pulses Help
~ non-modifiable copy of sound
0.2171
0
-0.1996
derived spectrogram
5000 Hz
0 Hz
modifiable TextGrid
words (85/85)
phones (297)
to be named in march it may be the most important
T IH B IY0 N EY1 M D IH0 N A A1 R CH IH1 M EY1 B IY0 D H AH M OW1 S T I M P
11.454434
Visible part 2.400257 seconds
13.854691
Total duration 25.309125 seconds
all in out sel bak
Group
29°C Mostly sunny
Search
ENG US
15:35 05-11-2025


5. TextGrid ISLE_SESS0131_BLOCKD02_01_sprt1
File Edit Time Play TextGrid Interval Boundary Tier Sound Analyses Spectrogram Pitch Intensity Formants Pulses Help
4.125000
0.9723
~ non-modifiable copy of sound
0.005951
-1
5000 Hz
derived spectrogram
190.9 Hz
modifiable TextGrid
1
i said white not bait
words (8/8)
2
A Y S E H D W AY 1 T N T EY1 T
phones (16)
0
Visible part 4.125000 seconds
4.125000
Total duration 4.125000 seconds
all in out sel bak
Group
Upcoming Earnings
Search
ENG US
15:37
05-11-2025


10. TextGrid ISLE_SESS0131_BLOCKD02_02_sprt1
File Edit Time Play TextGrid Interval Boundary Tier Sound Analyses Spectrogram Pitch Intensity Formants Pulses Help
2.190000
1.685000 (0.593 / s)
1
~ non-modifiable copy of sound
0
-0.634
5000 Hz
derived spectrogram
4187 Hz
0 Hz
modifiable TextGrid
1
i said new not no
words (8)
2
A Y S E H N Y UW1 N T N OW1
phones (15/15)
2.190000
1.685000
0
Visible part 3.875000 seconds
3.875000
Total duration 3.875000 seconds
all in out sel bak
Group
Upcoming Earnings
Search
ENG US
15:39
05-11-2025


11. TextGrid F2BJRLP2
File Edit Time Play TextGrid Interval Boundary Tier Sound Analyses Spectrogram Pitch Intensity Formants Pulses Help
step
14.323719
0.2112
0.03191
-0.2113
~ non-modifiable copy of sound
5000 Hz
0 Hz
derived spectrogram
modifiable TextGrid
1
for hennessy it was another step along a distinguished
words (39/83)
2
F HH EH AH0 S IY0 I T W AH Z AH N AH DH ER0 S T EH1 P L AO1 NG AH D IH S T IH IH SH
phones (337)
1.358427
1.358427
12.965291
12.965291
Visible part 2.716855 seconds
15.682146
12.965291
Total duration 28.647437 seconds
all in out sel bak
Group
29°C Mostly sunny
Search
ENG US
15:40 05-11-2025


14. TextGrid F2BJRLP3
File Edit Time Play TextGrid Interval Boundary Tier Sound Analyses Spectrogram Pitch Intensity Formants Pulses Help
legislative
15.000000
0.1828
0.01224
-0.1867
~ non-modifiable copy of sound
5000 Hz
0 Hz
derived spectrogram
modifiable TextGrid
1
and for last year's legislative approval of thirty
words (50/95)
2
AE1 N D F ER0 L AE1 S T Y I R Z L EH JH AH S EY2 T I V AH P R UW1 V L AH1 V TH
phones (385)
1.238495
1.238495
13.761505
13.761505
Visible part 2.476990 seconds
16.238495
14.468318
Total duration 30.706813 seconds
all in out sel bak
Group
29°C Mostly sunny
Search
ENG US
15:42 05-11-2025



Key Findings on Alignment Accuracy:

1. **Word Boundaries:** The start and end times for words were highly accurate. Boundaries consistently aligned perfectly with changes in the speech signal, such as the onset of a new word or the silence between words.
2. **Phoneme Boundaries:** Alignment at the sound level was also strong. For example, the precise moment a speaker transitions from a vowel sound (e.g., 'I' in 'pin') to a nasal sound (e.g., 'N' in 'pin') was correctly marked on the spectrogram.

Observed Issues and Mismatches:

1. **Fast Speech Challenges:** In segments where the speaker spoke very quickly or mumbled, MFA sometimes struggled with brief, unstressed sounds (like the vowel in "to"). This occasionally resulted in the phoneme being assigned a very short, potentially inaccurate duration.
2. **Acoustic Variability:** Since the acoustic model is generic (english_us_arpa), it occasionally misjudged the timing of words that had pronunciations slightly different from standard American English, leading to minor shifts in the phoneme boundaries.