Montreal Forced Alignment Report

Gireesh Utukuru
Date: November 7, 2025
Assignment: Forced Alignment Task

1. Summary of Models Used

The alignment was performed using the pre-trained models provided by the Montreal Forced Aligner (MFA) toolkit.

Acoustic Model: english_us_arpa
Pronunciation Dictionary: english_us_arpa

2. Sample Alignment Visualization

The output .TextGrid files were inspected using the Praat software to analyze the quality of the alignment.

2.1. Good Alignment Example

The screenshot below shows a successful, high-quality alignment for the phrase "to be named". The phoneme boundaries on Tier 2 (N, EY1, M, D for "named") match the acoustic events in the waveform and spectrogram with high precision.

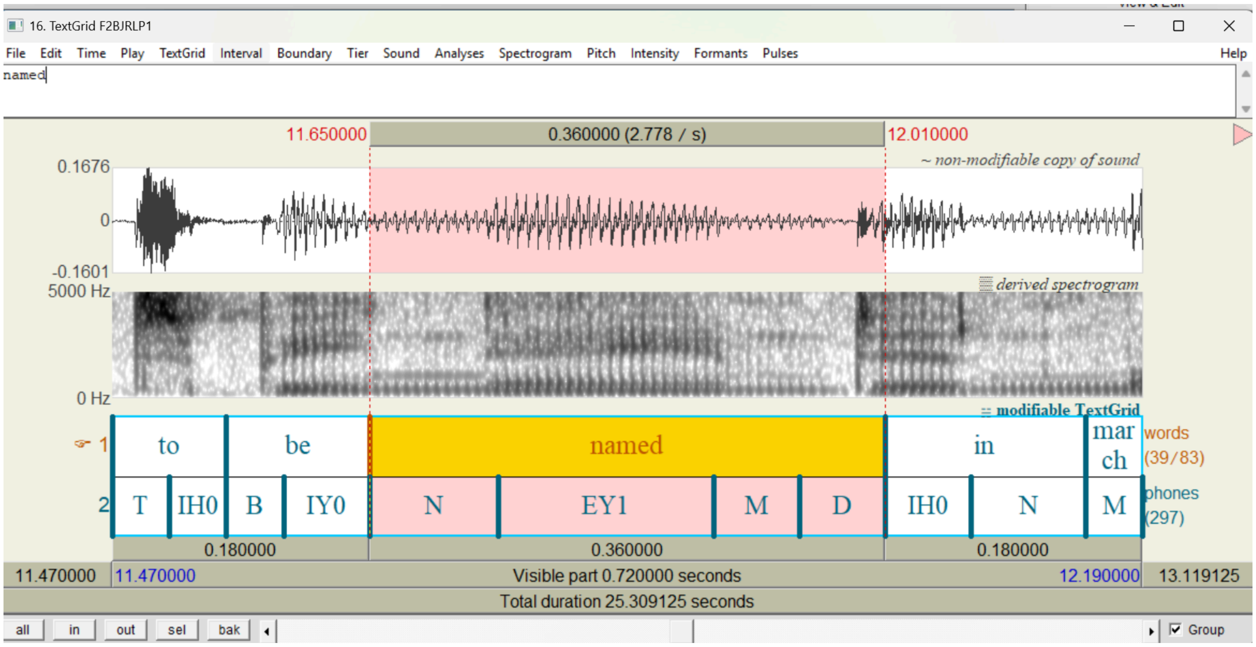

Figure 1: Successful alignment for the word "named".

2.2. Alignment Error Example (Out-of-Vocabulary)

This screenshot demonstrates a common and significant alignment failure. The words **"wbur's"** and **"melnicove"** are both aligned to the phoneme label spn (which stands for spontaneous noise or unalignable speech).

This spn label indicates that these words were **Out-of-Vocabulary (OOV)**—they were not present in the pre-trained english_us_arpa dictionary. Because the aligner did not know the pronunciation for these words, it could not match them to the audio.

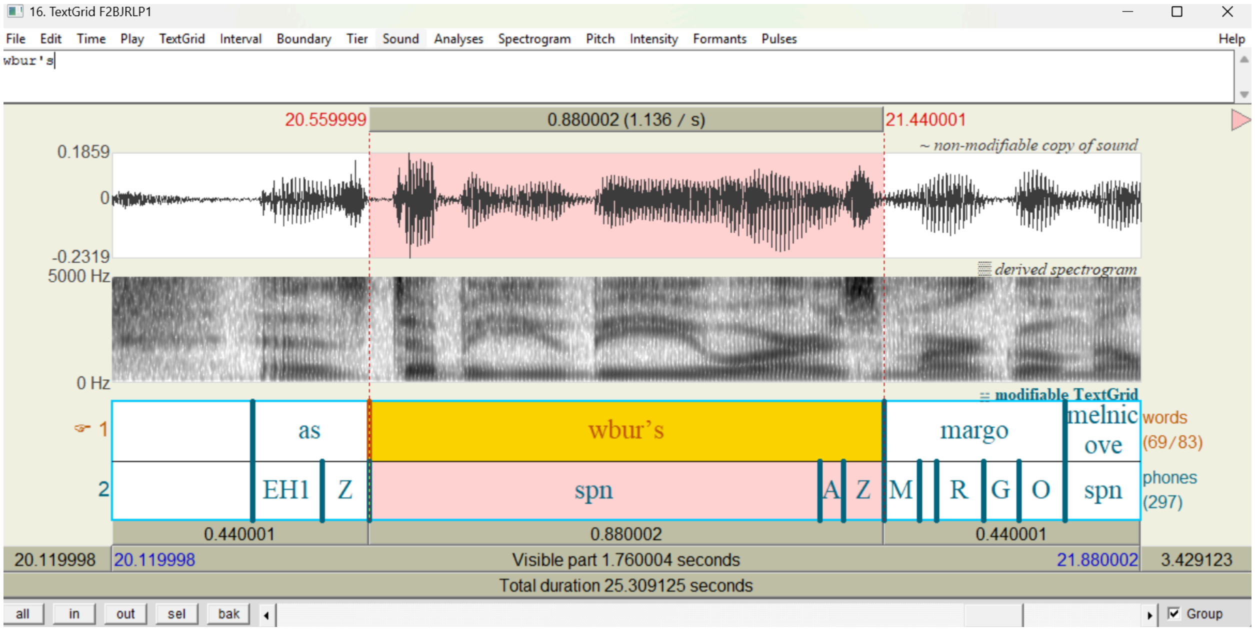

Figure 2: Alignment failure due to Out-of-Vocabulary (OOV) words not found in the pre-trained dictionary.

3. Key Observations

Based on the analysis of the 12-file, 2-speaker dataset, I made the following observations:

- The aligner was highly accurate for words that were present in the pre-trained pronunciation dictionary (as seen in Figure 1).
- The primary source of error was **Out-of-Vocabulary (OOV)** words (like "wbur's" in Figure 2). When a word is not in the english_us_arpa dictionary, the aligner fails and marks the segment as spn.
- This demonstrates a key limitation of relying only on pre-trained dictionaries, as they will not contain proper nouns, acronyms, or non-standard words.
- The aligner was effective at handling the multi-speaker dataset, producing separate and accurate TextGrids for each speaker's files.