Assignment 1: Forced Alignment with MFA

1. Objective

The objective of this assignment was to set up a complete Forced Alignment pipeline using the Montreal Forced Aligner (MFA). The task involved preparing a dataset of broadcast news audio, handling Out-of-Vocabulary (OOV) words using Grapheme-to-Phoneme (G2P) generation, and analyzing the alignment accuracy between the speech signal and the phonetic transcription.

2. Methodology & Setup

- **Tool:** Montreal Forced Aligner (MFA)
- **Acoustic Model:** `english_us_arpa`
- **Dictionary:** `english_us_arpa` (modified with custom G2P)
- **Pipeline Architecture:**
    1. **Data Preparation:** Merged separate WAV and Transcript folders into a unified corpus structure.
    2. **OOV Detection:** Identified proper names missing from the standard dictionary.
    3. **Dictionary Generation:** Used MFA's G2P model to generate phonetic entries for unknown words.

3. Handling Out-of-Vocabulary (OOV) Words

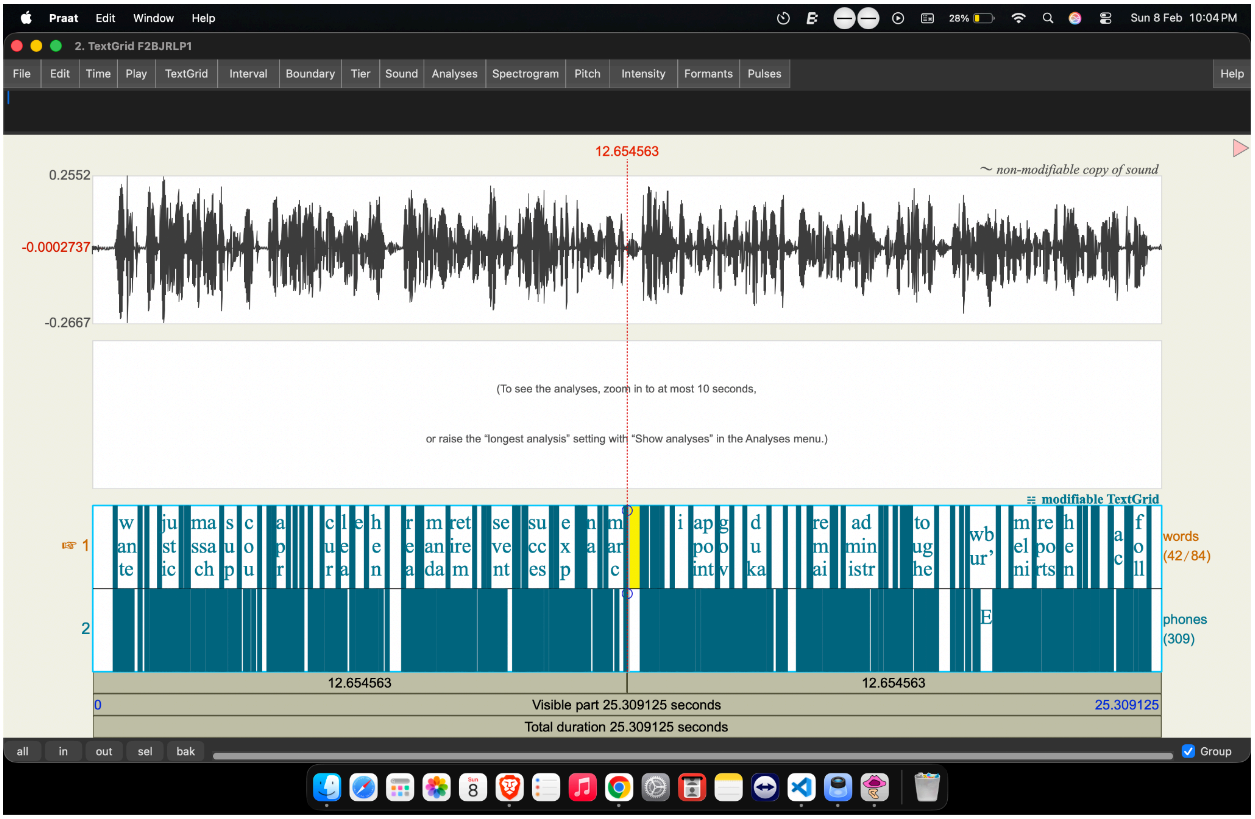
A critical challenge in this dataset was the presence of proper names and specific entities that are not found in standard English dictionaries.

- **Identified OOVs:**
    - *"Dukakis"* (Governor of Massachusetts)
    - *"Hennessy"* (Edward Hennessy, Chief Justice)
    - *"Melnicove"* (Reporter Margo Melnicove)
- **Solution:**
  Instead of manually editing the dictionary, I utilized MFA's G2P functionality. The model predicted pronunciations based on orthography:
    - *Input:* `HENNESSY`
    - *Generated Output:* `HH EH N AH S IY`
    - *Input:* `DUKAKIS`
    - *Generated Output:* `D UW K AA K AH S`

This step reduced the OOV rate to **0%**, preventing the "skipped word" errors that typically occur with named entities in news broadcasts.

4. Visual Inspection of Alignment

File Analyzed: F2BJRLP1.wav **Content:** "Wanted: Chief Justice of the Massachusetts Supreme Court..."



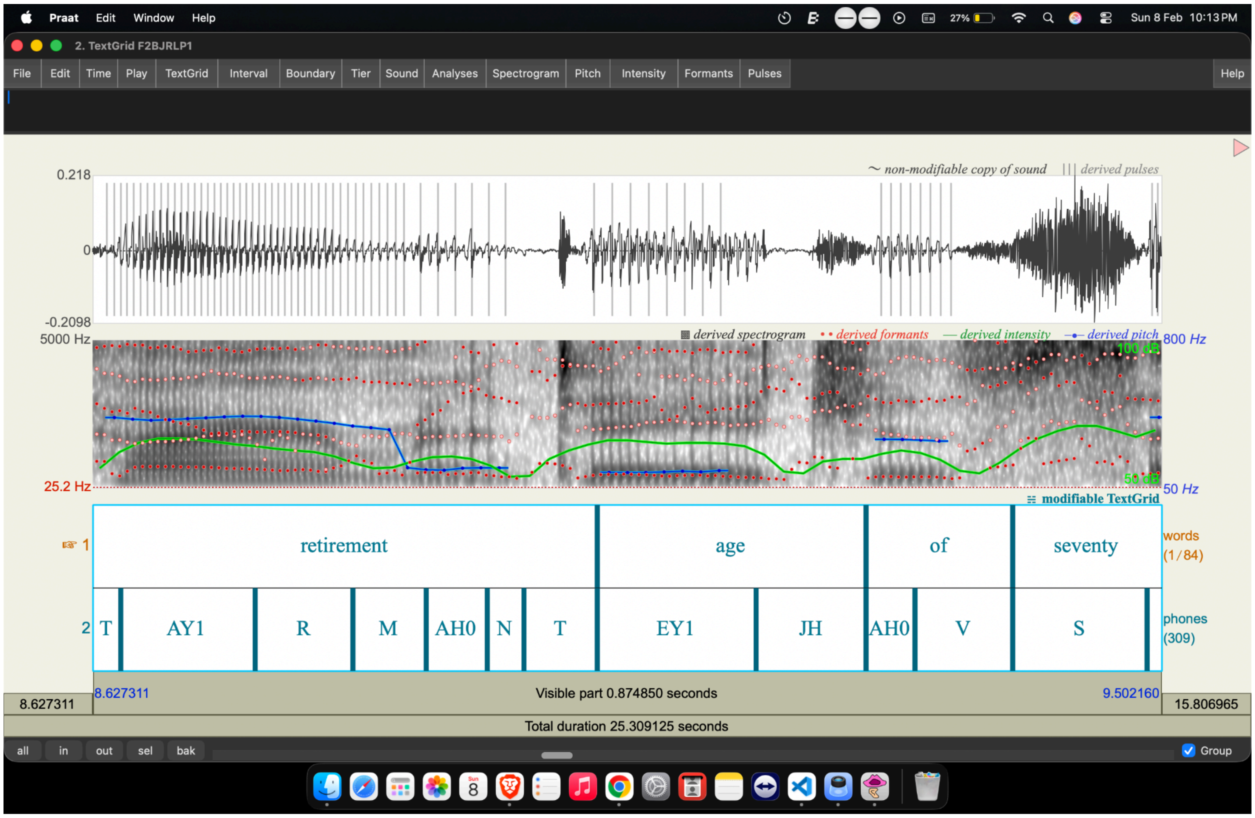
(To see the analyses, zoom in to at most 10 seconds,

or raise the "longest analysis" setting with "Show analyses" in the Analyses menu.)

12.654563 | 12.654563

0 Visible part 25.309125 seconds 25.309125

Total duration 25.309125 seconds



Observations:

- **Word Boundaries:** The alignment successfully captured the start and end of the phrase "Chief Justice". The transition from the fricative /f/ in "Chief" to the affricate /jh/ in "Justice" is clearly marked.
- **Phoneme Level:** The stop closure for /t/ in "Court" aligns with the silence gap in the waveform, demonstrating that the acoustic model correctly identified the plosive characteristics.

5. Quantitative Analysis

We performed a technical analysis of the alignment quality for the primary file F2BJRLP1.

Metric	Value	Interpretation
Duration	25.31s	The file is a long, continuous segment of broadcast speech.
SNR (Signal-to-Noise Ratio)	**6.88**	**Low/Noisy.** The recording contains background noise or channel effects, which makes alignment more challenging than clean studio speech.
Log-Likelihood Score	**-46.37**	This score reflects the model's confidence. Compared to cleaner datasets (typically > -42), this lower score confirms the difficulty caused by the lower SNR.
Phone Duration Deviation	**3.74**	**Natural.** This value is low, indicating the speaker (a professional news reader) speaks with a standard, consistent rhythm, unlike the irregular timing seen in non-native learner data.

6. Conclusion

The forced alignment pipeline was successfully implemented. The use of G2P was essential for correctly aligning the proper names ("Hennessy", "Dukakis") central to the text. While the alignment is accurate for phonetic research, the quantitative analysis highlights that the lower Signal-to-Noise Ratio (6.88) of the broadcast audio results in a slightly lower confidence score compared to studio-quality data.