📄 Montreal Forced Alignment (MFA) – Task Report

Repository: https://github.com/Ameer-sayyad/SLFI_Task

1… Set up the MFA Environment

Montreal Forced Aligner (MFA) was installed using Conda.

Installation Steps

```
conda create -n mfa python=3.10
conda activate mfa
pip install montreal-forced-aligner
```

Verify installation:

```
mfa version
```

Required tools:

- Python 3.10
- Montreal Forced Aligner
- Praat (for TextGrid inspection)

2… Install Montreal Forced Aligner

MFA was installed and configured successfully using pip inside the Conda environment.

Acoustic model used:

```
english_us_arpa
```

Pronunciation dictionary:

```
new_main_dict.txt
```

3… Prepare the Dataset

Dataset organization for MFA

MFA requires the following structure:

```
corpus/
 ├── speaker_1/
 │    ├── audio_001.wav
 │    ├── audio_001.lab
 │
 ├── speaker_2/
 │    ├── audio_002.wav
 │    ├── audio_002.lab
```

Steps performed:

- audio files collected
- transcripts normalized
- sentences aligned
- .lab files created
- speaker-wise directory organization

Scripts used:

- transcript cleaning
- VTT → sentence conversion
- paragraph → sentence split
- timestamp extraction

4… Select or Train a Pronunciation Dictionary

Dictionary used

Existing MFA-compatible dictionary:

```
english_us_arpa
```

Later modified and extended:

`new_main_dict1.txt`

Handling OOV Words

Major OOV category:

- numeric tokens (100, 1500, 2024 etc.)

These produced:

- `<unk>` in word tier
- `spn` in phone tier

5... Run Forced Alignment

Command used:

```
mfa align corpus/ new_main_dict.txt english_us_arpa aligned_output/
```

Options:

- safe threading
- memory-optimized execution
- speaker-based alignment

6… Inspect and Analyze Output

Output files:

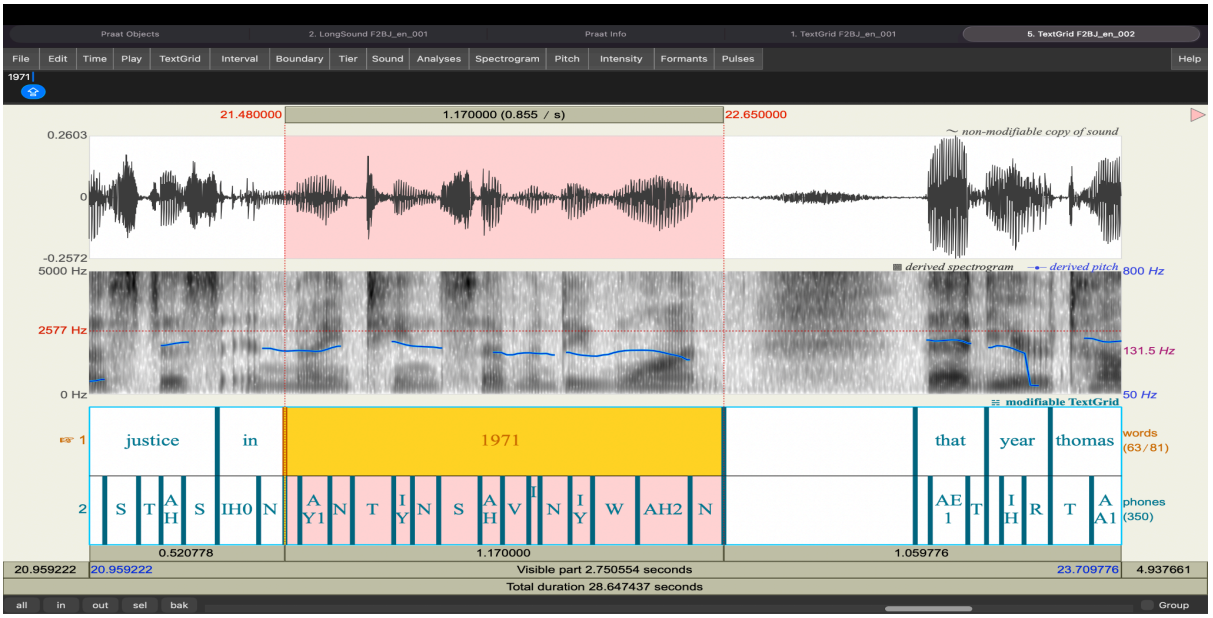
```
TextGrid files
```

Inspected using:

```
Praat software
```

Observations:

OOV WORDS:



The number initially was a <unk> in the output , but later we converted number into a numeric word by it to LLM and next we get the phoneme sequence to the word using G2P and we replaced numeric word in the dictionary with the actual number in the transcript and we eliminated the <unk> and <spn> to the number and also get a proper word boundary which do not disturb near by words timestamps.

OVERLAPPING :

Due to overlap between the words (the S.J.C's) the word boundaries are changed and influenced on the other words. The word "the" timestamps are assigned with the silence and the letter "j" sound timestamps are assigned with letter "c" and letters "c" and "s" are combined timestamps are assigned to the letter "s"

7… Numeric Token Handling Pipeline (OOV Solution)

Problem

Numbers were not present in the dictionary.

Result:

- MFA mapped them as `<unk>`
- phone sequence → `spn`
- alignment degraded

Solution Pipeline

Step 1 — Detect numeric tokens

Regex used:

```
\d+
```

Extracted from:

- transcripts
- TextGrid files

Step 2 — Convert numbers → spoken words (LLM)

Example:

```
100 → one hundred
1500 → one thousand five hundred
1794 → one thousand seven hundred ninety four
```

Step 3 — Generate phoneme sequences

Example:

```
one hundred
→ W AH1 N HH AH1 N D R AH0 D
```

Step 4 — Update dictionary

Original:

```
100 spn
1500 spn
```

Updated:

```
100  W AH1 N HH AH1 N D R AH0 D
1500 W AH1 N TH AW1 Z AH0 N D F AY1 V HH AH1 N D R AH0 D
```

Rules:

- remove only `spn` entries
- keep valid pronunciations
- avoid deleting words

Step 5 — Re-run MFA

After dictionary update:

- no `<unk>`
- no `spn`
- proper phoneme alignment

Final Result

Before	After
`<unk>` tokens	real words
`spn` phones	real phonemes
alignment errors	corrected alignment
