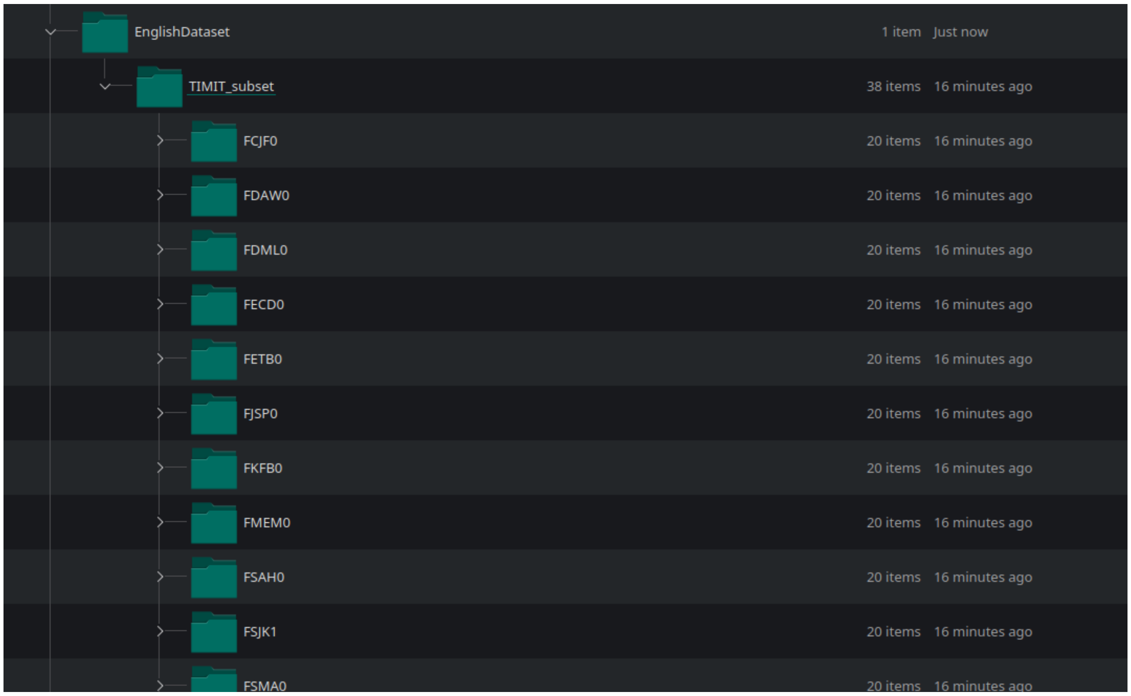
Use Case 3: English alignments

We will align English data in this tutorial. The dataset we use is a subset of the DARPA TIMIT Acoustic-Phonetic Continuous Speech database (Garofolo et al. 1993). The full corpus can be obtained here: https://www.kaggle.com/datasets/mfekadu/darpa-timit-acousticphonetic-continuous-speech .

1) Create a corpus folder and extract the data

You can either directly etract the TIMIT subset somewhere on your computer, or create a new folder that will contain the raw data and the alignments. For this use case, I generated a folder called "EnglishDataset" and deployed the data there. The folder structure should look like this, with several subdirectories for each speaker:



EnglishDataset	1 item	Just now
TIMIT_subset	38 items	16 minutes ago
FCJF0	20 items	16 minutes ago
FDAW0	20 items	16 minutes ago
FDML0	20 items	16 minutes ago
FECD0	20 items	16 minutes ago
FETB0	20 items	16 minutes ago
FJSP0	20 items	16 minutes ago
FKFB0	20 items	16 minutes ago
FMEM0	20 items	16 minutes ago
FSAH0	20 items	16 minutes ago
FSJK1	20 items	16 minutes ago
FSMA0	20 items	16 minutes ago

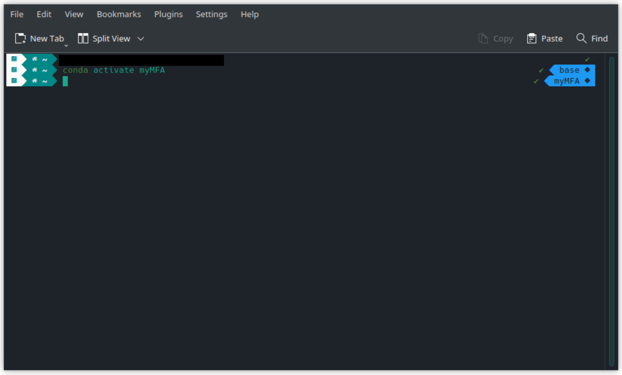
This subset is already prepared for the alignment in the MFA, meaning that the .txt files containing the utterances where already transformed into .lab files. Note that you can use also the respective .txt files in the original corpus, but you have to replace the .txt suffix with .lab and you have to remove the time stemps within the files.

2) Open Anaconda and activate your MFA environment

Open your terminal/console/command line and activate the environment in which the MFA installed:

[ ]:

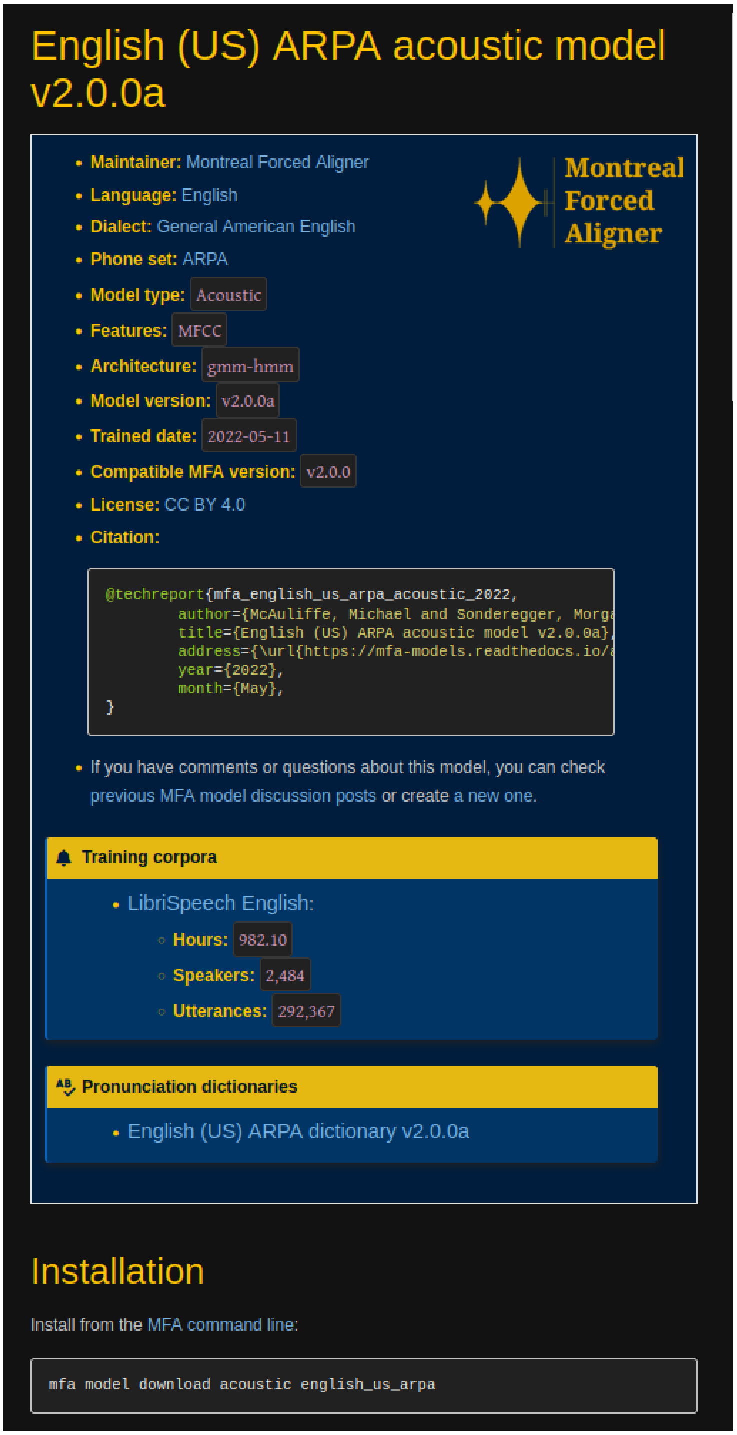
```
conda activate myMFA
```



```
~ [redacted]                                          ✓
~ conda activate myMFA                         ✓ base
~                                              ✓ myMFA
```

3) Download the acoustic model and dictionary for English

In this step you have to download the acoustic model and dictionary for the English alignments. These can be found here: https://mfa-models.readthedocs.io/en/latest/ . If you search for English, you will find two different acoustic models. You will see that these models differ by the feature extraction and amount of data that were used for training.



English (US) ARPA acoustic model v2.0.0a

Montreal
Forced
Aligner

- **Maintainer:** Montreal Forced Aligner
- **Language:** English
- **Dialect:** General American English
- **Phone set:** ARPA
- **Model type:** Acoustic
- **Features:** MFCC
- **Architecture:** gmm-hmm
- **Model version:** v2.0.0a
- **Trained date:** 2022-05-11
- **Compatible MFA version:** v2.0.0
- **License:** CC BY 4.0
- **Citation:**

```
@techreport{mfa_english_us_arpa_acoustic_2022,
        author={McAuliffe, Michael and Sonderegger, Morga
        title={English (US) ARPA acoustic model v2.0.0a},
        address={\url{https://mfa-models.readthedocs.io/a
        year={2022},
        month={May},
}
```

- If you have comments or questions about this model, you can check previous MFA model discussion posts or create a new one.

Training corpora

- LibriSpeech English:
  - **Hours:** 982.10
  - **Speakers:** 2,484
  - **Utterances:** 292,367

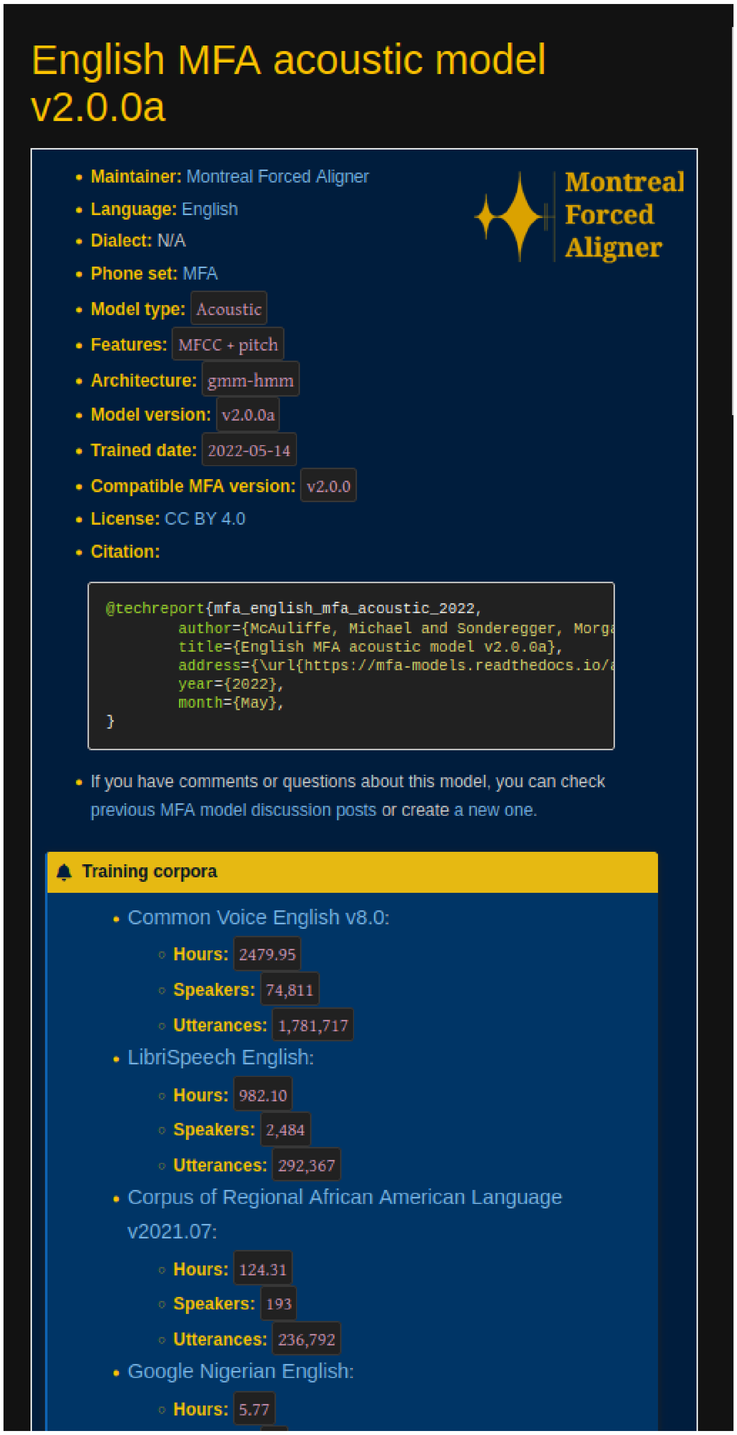
Pronunciation dictionaries

- English (US) ARPA dictionary v2.0.0a

Installation

Install from the MFA command line:

```
mfa model download acoustic english_us_arpa
```



English MFA acoustic model v2.0.0a

Montreal Forced Aligner

- **Maintainer:** Montreal Forced Aligner
- **Language:** English
- **Dialect:** N/A
- **Phone set:** MFA
- **Model type:** Acoustic
- **Features:** MFCC + pitch
- **Architecture:** gmm-hmm
- **Model version:** v2.0.0a
- **Trained date:** 2022-05-14
- **Compatible MFA version:** v2.0.0
- **License:** CC BY 4.0
- **Citation:**

```
@techreport{mfa_english_mfa_acoustic_2022,
        author={McAuliffe, Michael and Sonderegger, Morga
        title={English MFA acoustic model v2.0.0a},
        address={\url{https://mfa-models.readthedocs.io/
        year={2022},
        month={May},
}
```

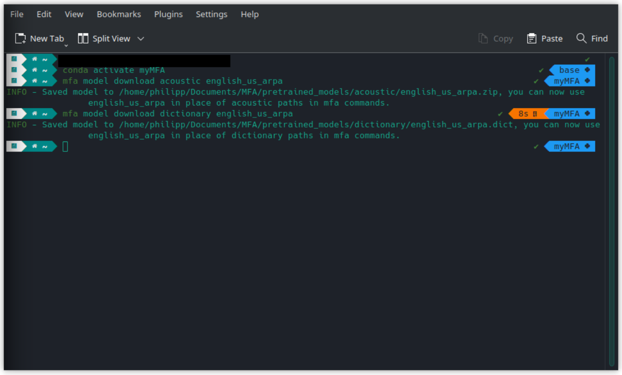
- If you have comments or questions about this model, you can check previous MFA model discussion posts or create a new one.

Training corpora

- Common Voice English v8.0:
  - **Hours:** 2479.95
  - **Speakers:** 74,811
  - **Utterances:** 1,781,717
- LibriSpeech English:
  - **Hours:** 982.10
  - **Speakers:** 2,484
  - **Utterances:** 292,367
- Corpus of Regional African American Language v2021.07:
  - **Hours:** 124.31
  - **Speakers:** 193
  - **Utterances:** 236,792
- Google Nigerian English:
  - **Hours:** 5.77

These two models are also accompanied by different dictionaries. We fill first download the ARPA acoustic model and dictionary:

```
[ ]: mfa model download acoustic english_us_arpa
     mfa model download dictionary english_us_arpa
```



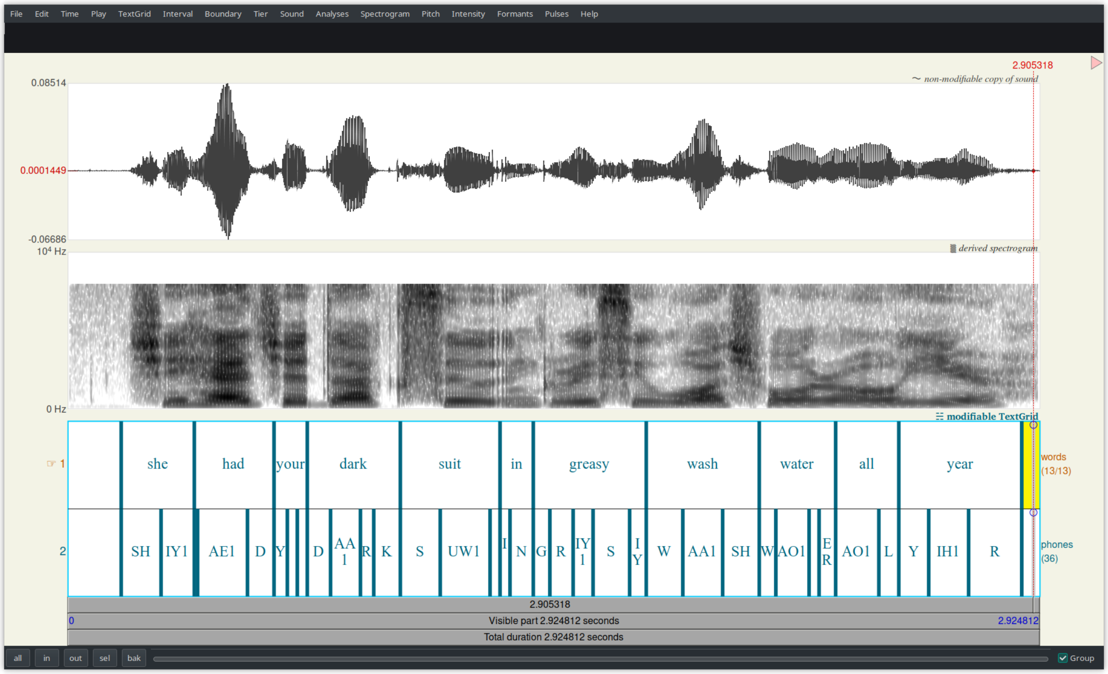
```
~ [redacted]                                                                  ✓
~ conda activate myMFA                                                  ✓ base
~ mfa model download acoustic english_us_arpa                          ✓ myMFA
INFO - Saved model to /home/philipp/Documents/MFA/pretrained_models/acoustic/english_us_arpa.zip, you can now use
           english_us_arpa in place of acoustic paths in mfa commands.
~ mfa model download dictionary english_us_arpa                    ✓ 8s  myMFA
INFO - Saved model to /home/philipp/Documents/MFA/pretrained_models/dictionary/english_us_arpa.dict, you can now use
           english_us_arpa in place of dictionary paths in mfa commands.
~                                                                      ✓ myMFA
```

4) Align the English data

Since we are using the ARPA model, enter the following command to generate the alignments. Adapt the paths, if necessary.

```
[ ]: mfa align --clean ~/Documents/EnglishDataset/TIMIT_subset english_us_arpa␣
     ↪english_us_arpa ~/Documents/EnglishDataset/ARPA_alignments/
```

After a successfull run, your should see the alignments in the ARPA_alignments folder. Inspect the TextGrids in Praat.

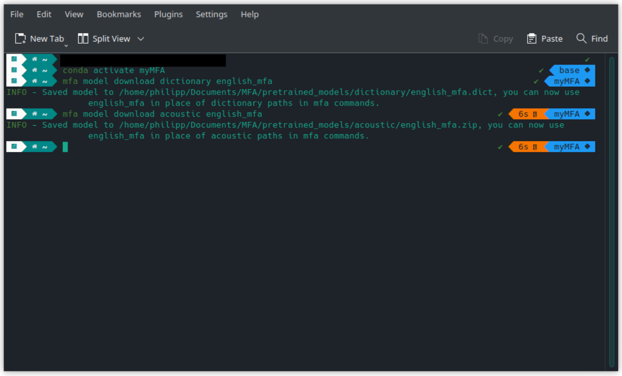


You will see that the phone labels are not in IPA as before, but in the ARPABET (https://en.wikipedia.org/wiki/ARPABET). Lets compare the labels and alignments with the second dictionary from the MFA.

5) Download the English MFA acoustic model and dictionary

Use the following commands to download the acoustic model and dictionary of the MFA:

```
[ ]: mfa model download acoustic english_mfa
     mfa model download dictionary english_mfa
```



```
~ [redacted]
~ conda activate myMFA
~ mfa model download dictionary english_mfa
INFO - Saved model to /home/philipp/Documents/MFA/pretrained_models/dictionary/english_mfa.dict, you can now use
        english_mfa in place of dictionary paths in mfa commands.
~ mfa model download acoustic english_mfa
INFO - Saved model to /home/philipp/Documents/MFA/pretrained_models/acoustic/english_mfa.zip, you can now use
        english_mfa in place of acoustic paths in mfa commands.
~
```

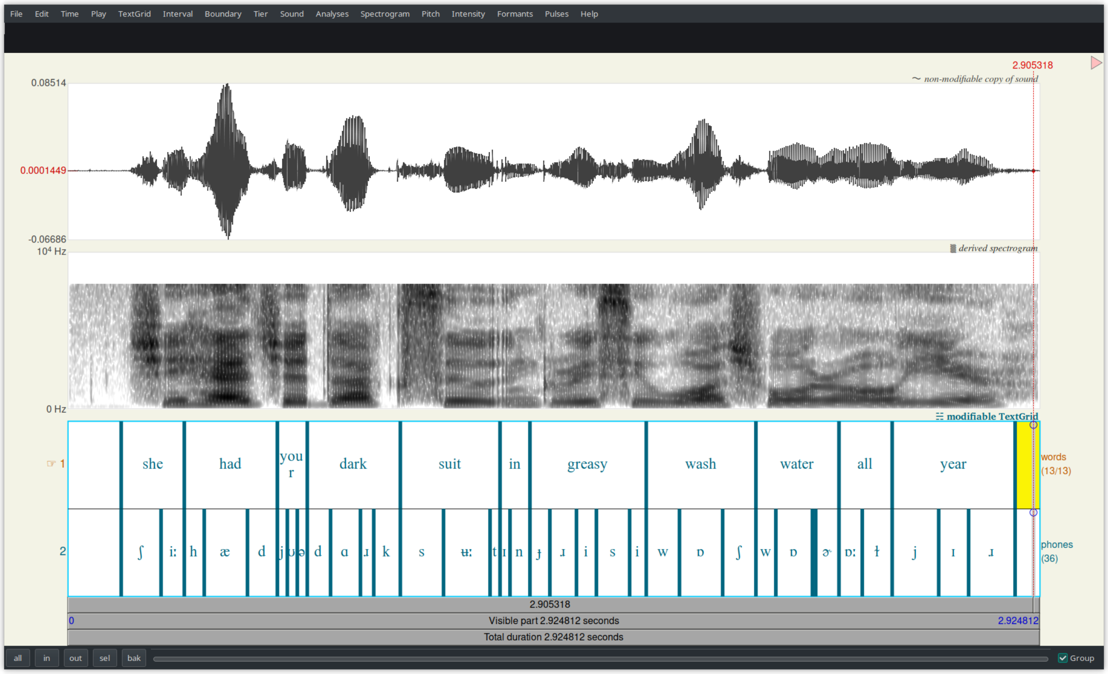
By inspecting the dictionaries under ~/Documents/MFA/pretrained_models/dictionary/ (Windows: C:/Users/username/Documents/MFA/pretrained_models/dictionary/), it can be seen that they differ not only in the transcription standard (ARPA: ARPABET, MFA: IPA) but also to the number of entries (ARPA: 206.366, MFA: 200.695).

5) Align the English data using the MFA model

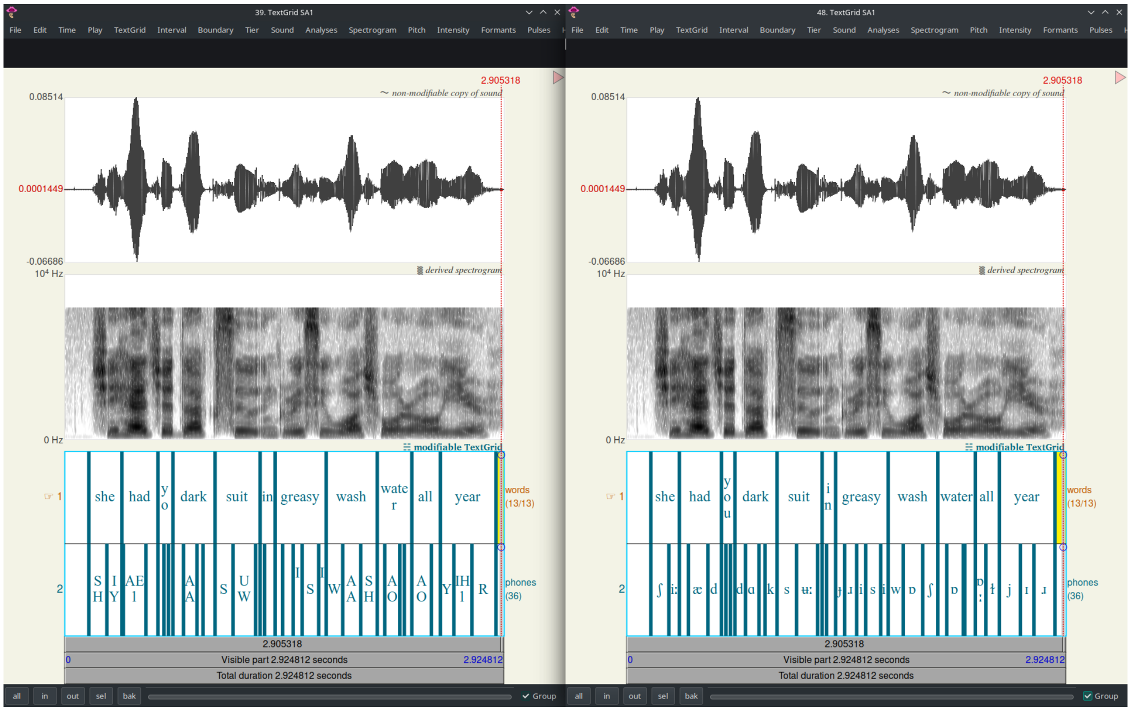
In order to align the data with the MFA model, use the code above, but replace “english_us_arpa” with “english_mfa” for both the dictionary and the acoustic model:

```
[ ]: mfa align --clean ~/Documents/EnglishDataset/TIMIT_subset english_mfa␣
      ↪english_mfa ~/Documents/EnglishDataset/MFA_alignments/
```

Under MFA_alignments you should find the TextGrids to use them in Praat:


File Edit Time Play TextGrid Interval Boundary Tier Sound Analyses Spectrogram Pitch Intensity Formants Pulses Help
2.905318
~ non-modifiable copy of sound
0.08514
0.0001449
-0.06686
10⁴ Hz
derived spectrogram
0 Hz
modifiable TextGrid
1
she had you r dark suit in greasy wash water all year
words (13/13)
2
ʃ iː h æ d j ʊ ɚ d ɑ ɹ k s ʉː t ɪ n ɟ ɹ i s i w ɒ ʃ w ɒ ɚ ɒː ɫ j ɪ ɹ
phones (36)
2.905318
0
Visible part 2.924812 seconds
2.924812
Total duration 2.924812 seconds
all in out sel bak
Group

You can now also compare the alignments of the two models:


39. TextGrid SA1
48. TextGrid SA1
File Edit Time Play TextGrid Interval Boundary Tier Sound Analyses Spectrogram Pitch Intensity Formants Pulses
2.905318
~ non-modifiable copy of sound
0.08514
0.0001449
-0.06686
10⁴ Hz
derived spectrogram
0 Hz
modifiable TextGrid
1
she had yo dark suit in greasy wash water all year
words (13/13)
2
SH IY AE l AA S UW IH S W AA SH AO AO Y IH R
phones (36)
she had you dark suit in greasy wash water all year
words (13/13)
ʃ iː æ d d ɑ k s ʉː ɟ ɹ i s i w ɒ ʃ ɒ ɒː ɫ j ɪ ɹ
phones (36)
2.905318
0
Visible part 2.924812 seconds
2.924812
Total duration 2.924812 seconds
all in out sel bak
Group

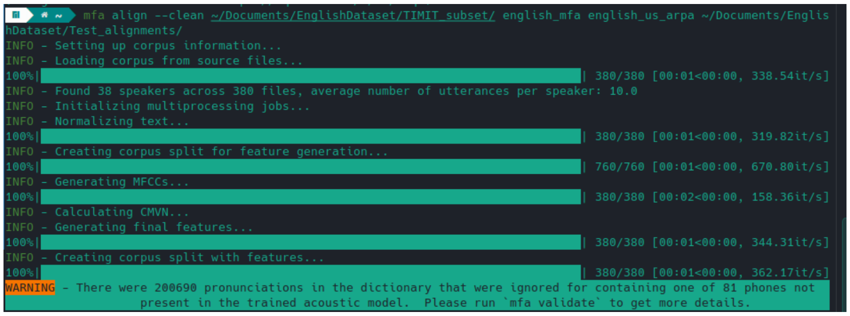
There should not be too much difference between the segmentation. However, you may prefer one acoustic model over the other.

Final remark

We saw that for a language like English we can have multiple acoustic models and dictionaries. Two points are noteworthy: First, if you plan to extend a dictionary manually, ensure to use the transcription standard for the phones. Otherwise the respective word will note be recognized. Second, do not use an acoustic model and another dictionary that was not used for training. For example, if we want to align the English data with the English MFA dictionary but with the ARPA acoustic model:

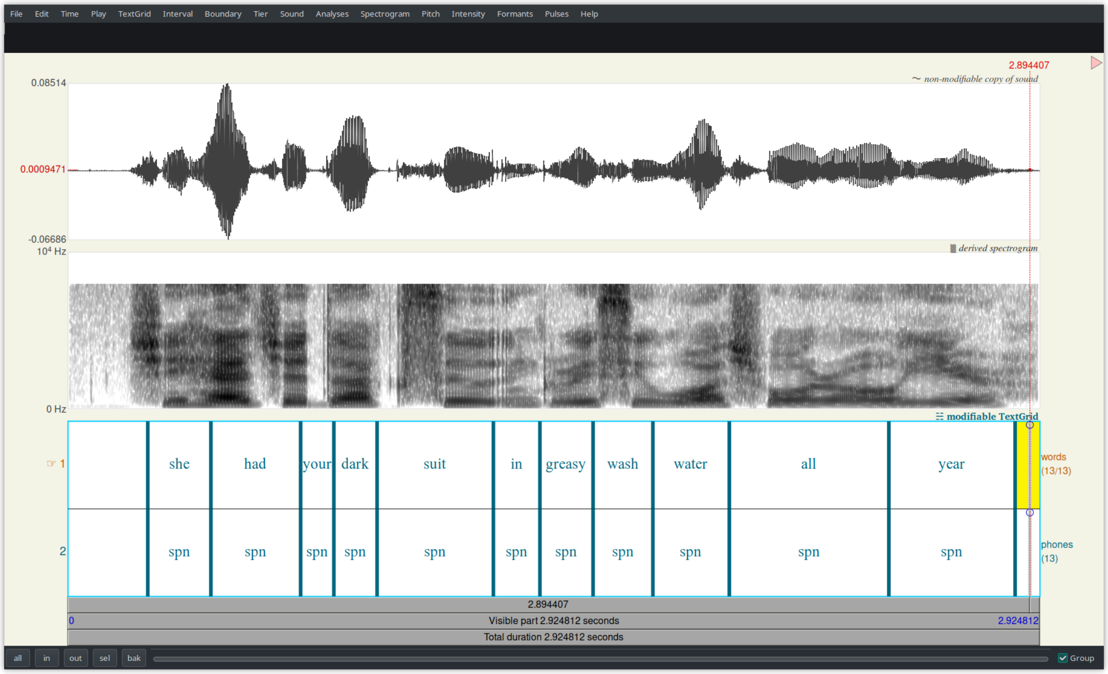
```
[ ]: mfa align --clean ~/Documents/EnglishDataset/TIMIT_subset english_mfa␣
     ↪english_us_arpa ~/Documents/EnglishDataset/MFA_alignments/
```

you will see this error message:



```
mfa align --clean ~/Documents/EnglishDataset/TIMIT_subset/ english_mfa english_us_arpa ~/Documents/Englis
hDataset/Test_alignments/
INFO - Setting up corpus information...
INFO - Loading corpus from source files...
100%|████████████████████████████████████████████████████| 380/380 [00:01<00:00, 338.54it/s]
INFO - Found 38 speakers across 380 files, average number of utterances per speaker: 10.0
INFO - Initializing multiprocessing jobs...
INFO - Normalizing text...
100%|████████████████████████████████████████████████████| 380/380 [00:01<00:00, 319.82it/s]
INFO - Creating corpus split for feature generation...
100%|████████████████████████████████████████████████████| 760/760 [00:01<00:00, 670.80it/s]
INFO - Generating MFCCs...
100%|████████████████████████████████████████████████████| 380/380 [00:02<00:00, 158.36it/s]
INFO - Calculating CMVN...
INFO - Generating final features...
100%|████████████████████████████████████████████████████| 380/380 [00:01<00:00, 344.31it/s]
INFO - Creating corpus split with features...
100%|████████████████████████████████████████████████████| 380/380 [00:01<00:00, 362.17it/s]
WARNING - There were 200690 pronunciations in the dictionary that were ignored for containing one of 81 phones not
             present in the trained acoustic model.  Please run `mfa validate` to get more details.
```

The reason is, that the phoneset of the MFA dictionary is different from the one that was used for the ARPA acoustic model. The MFA is thus not able to recognize the specific phones of the transcriptions of the words. This leads to TextGrids containing mis-alignments in the word tier and "spn" labels (default label to model unknown words) in the phone tier.



References

Garofolo, John S., et al. TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. Web Download. Philadelphia: Linguistic Data Consortium, 1993