Sprint 1
-
Multilingual Subtitle System

Nimrod Eshel

Introduction to Multilingual Subtitle System

Problem Area:

- The global demand for content accessibility in multiple languages.
- Challenges in creating accurate and synchronized subtitles for multilingual audiences.
- Impact on non-native speakers, hearing-impaired individuals, and content creators.

Opportunity:

- Develop a system that can generate accurate multilingual subtitles using data science and machine learning.

Vision for Multilingual Subtitle System

Data Science Solution:

- Use audio feature extraction (MFCCs, Chroma, Spectral Contrast) to analyze speech patterns.
- Implement machine learning models to recognize languages and generate subtitles.
- Utilize a robust dataset containing diverse languages and dialects.
    - Current focus on 4 languages (Arabic, English, Chinese, Spanish)

Technologies and Methods:

- Audio processing with librosa.
- Machine learning models for language recognition.
- Feature engineering and data preprocessing.

Potential Impact of the Solution

Accessibility:

- Enhance accessibility for non-native speakers to watch media from other languages.
- Broaden audience reach for content creators & entertainment companies.

Quality of Life:

- Improve learning experiences for language learners.
- Increase the quality and accuracy of subtitles for international audiences.

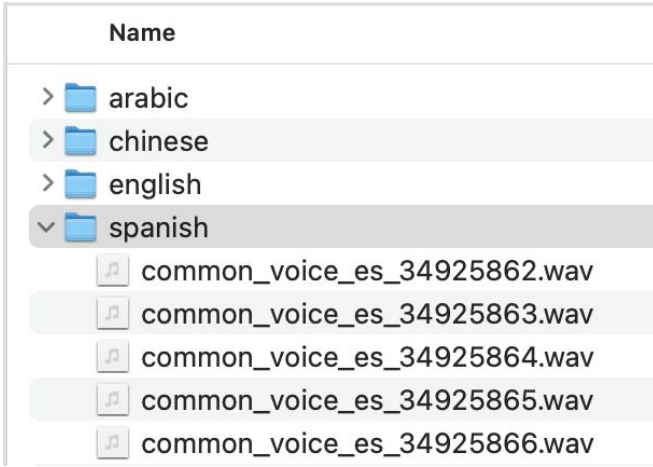
Market Expansion:

- Enable content creators to enter new markets by providing multilingual support.
- Support diverse cultural representation in media.
- Diversify audience by presenting more language options to watch content in.

Dataset and Preliminary EDA Findings

Data Set:

- Extracted languages from Mozilla Common Voice
- Converted to .wav files
- Extracted MFCCs, Chroma, Spectral Contrast
  - MFCCs: Columns 0 to 39
  - Delta Derivative MFCC: Columns 40 to 79
  - Delta-squared MFCC: Columns 80 to 119
  - Chroma Features: Columns 120 to 131
  - Spectral Contrast: Columns 132 to 138

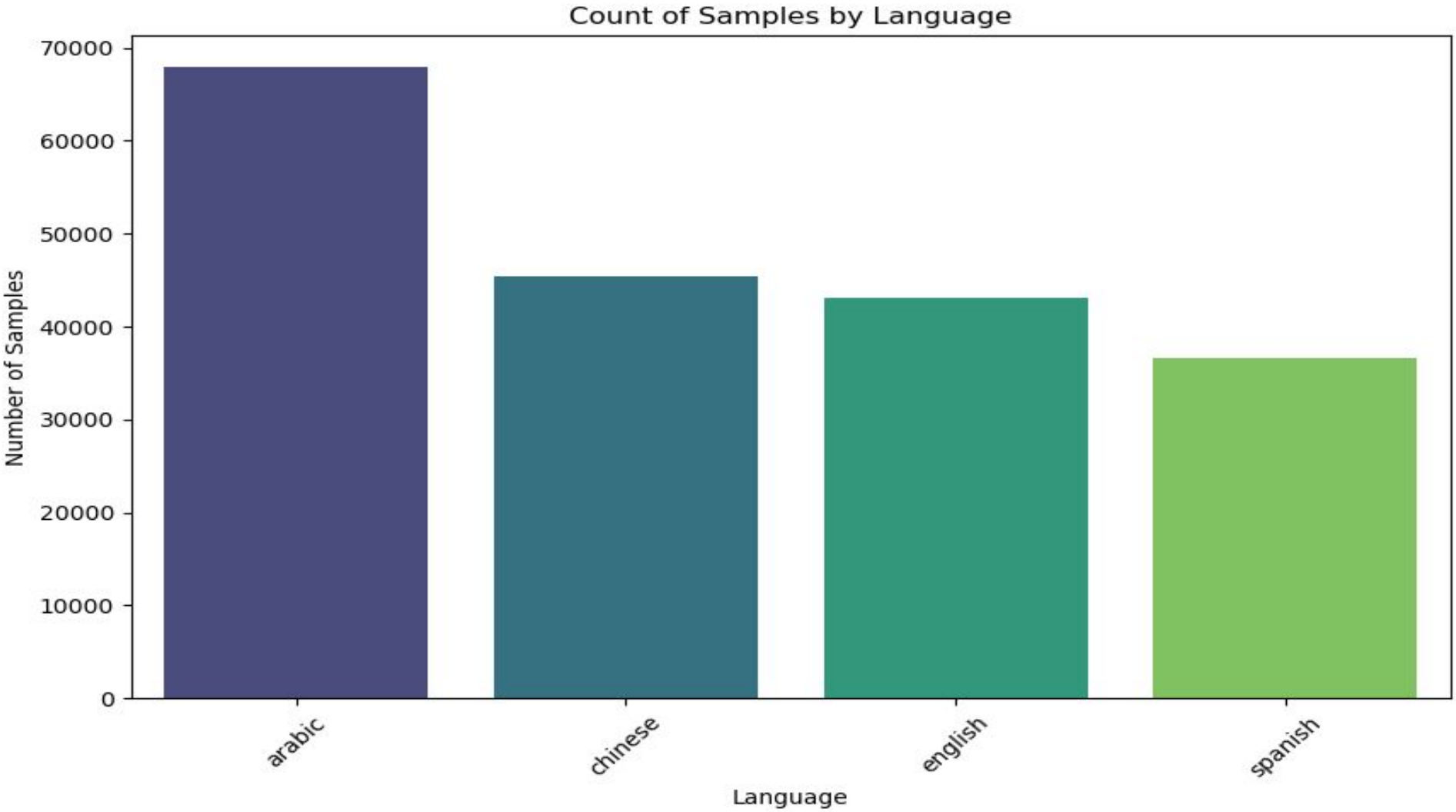


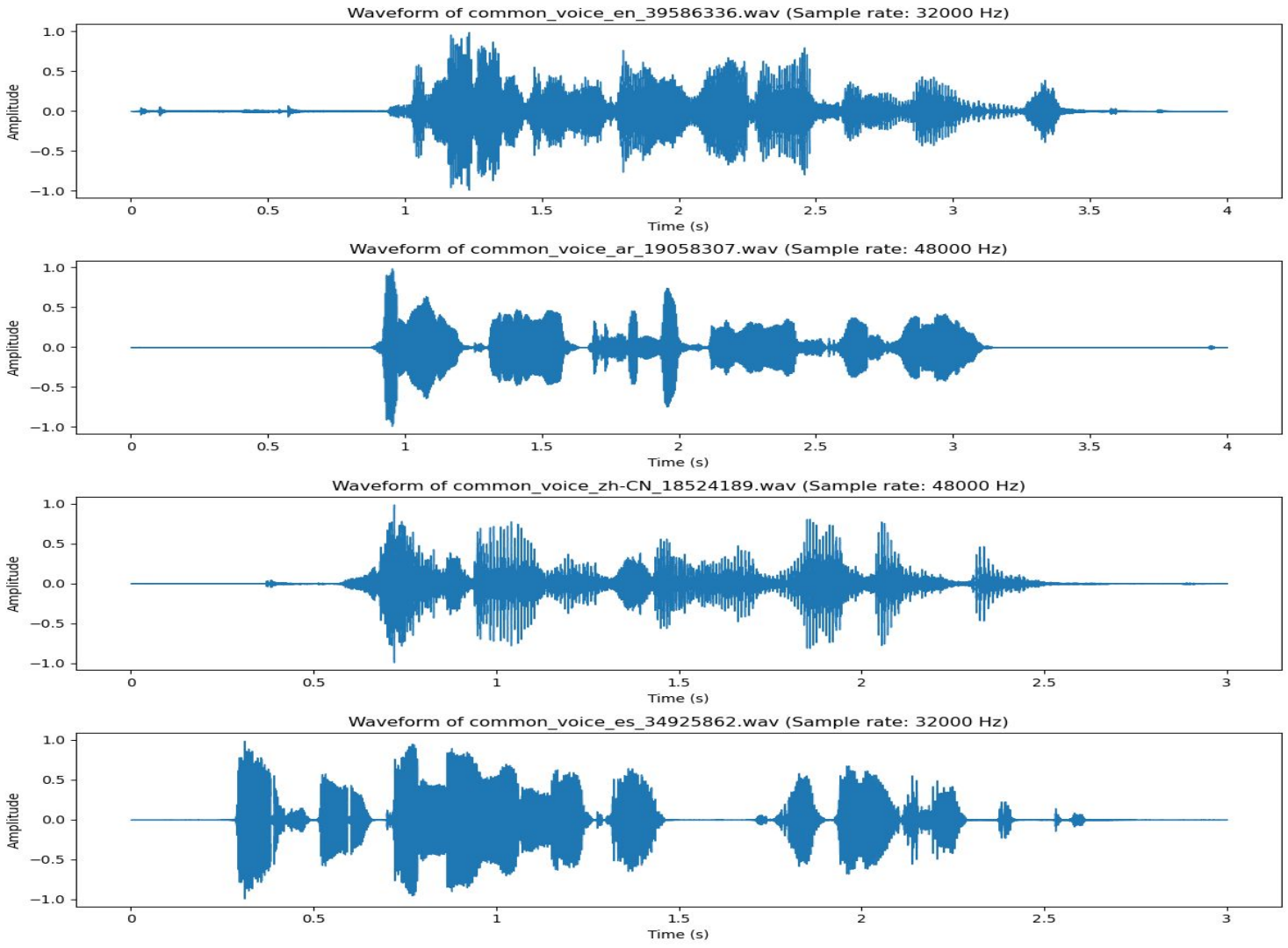
7	8	9	...	131	132	133	134	135	136	137	138	language	filename
.715307	-7.477776	-12.611868	...	0.260017	29.080616	16.073751	21.750985	24.338790	26.617756	46.909442	43.771037	english	common_voice_en_39594868.wav
898661	-16.897488	-0.571906	...	0.251122	23.249014	16.358835	23.411673	25.204130	28.218494	40.597539	44.491157	english	common_voice_en_39747444.wav
237065	-16.786737	6.401083	...	0.309458	25.558249	18.043898	22.215089	28.875768	41.785869	39.828661	50.301511	english	common_voice_en_39605105.wav
.513179	-0.274636	-2.809067	...	0.314104	20.875108	22.168876	25.364991	26.356457	36.447688	46.463369	46.004927	english	common_voice_en_39746982.wav
325612	-17.233961	-8.247074	...	0.391886	22.819066	13.254295	19.408762	31.489025	30.014653	34.741143	34.860893	english	common_voice_en_39806983.wav

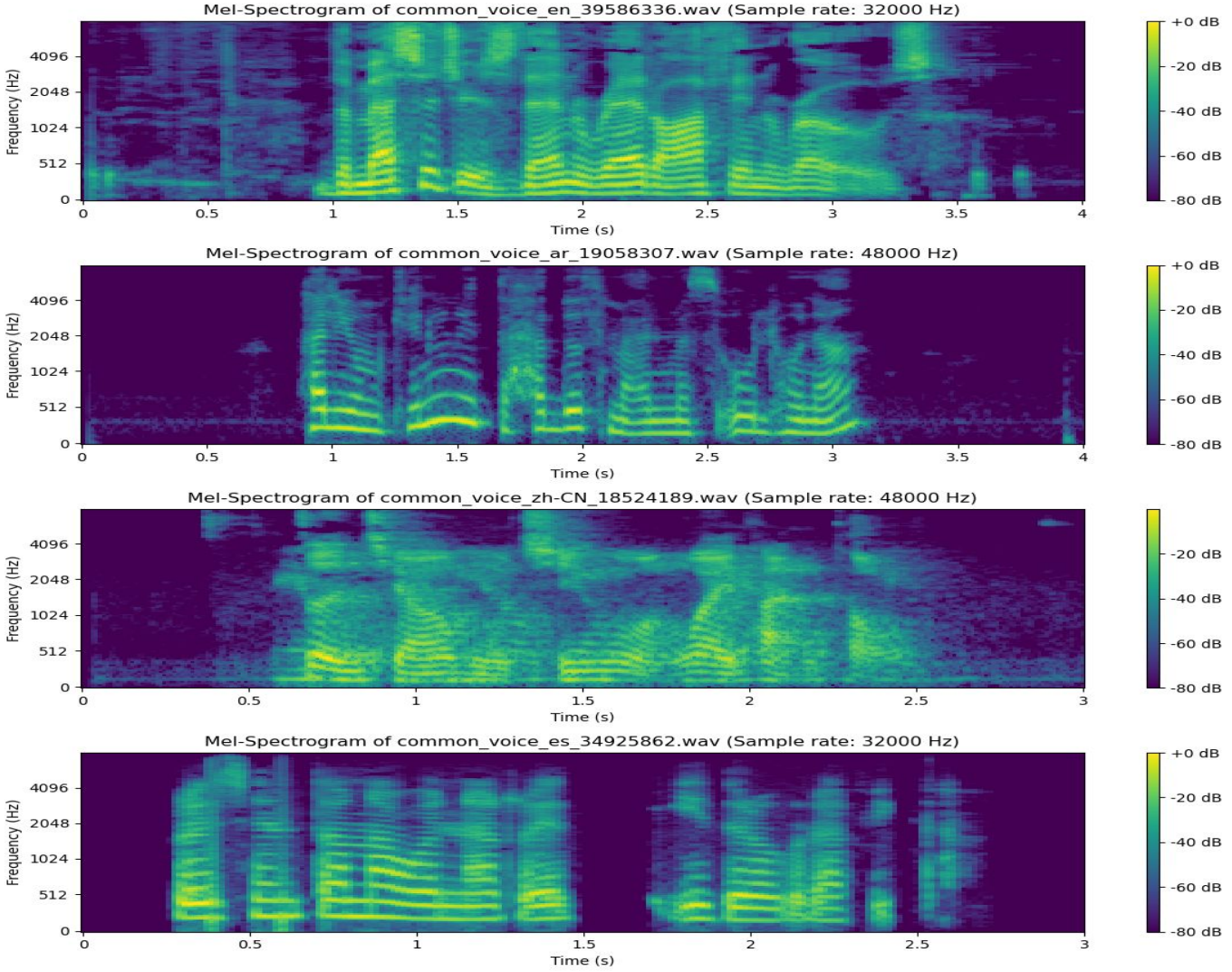
Dataset and Preliminary EDA Findings

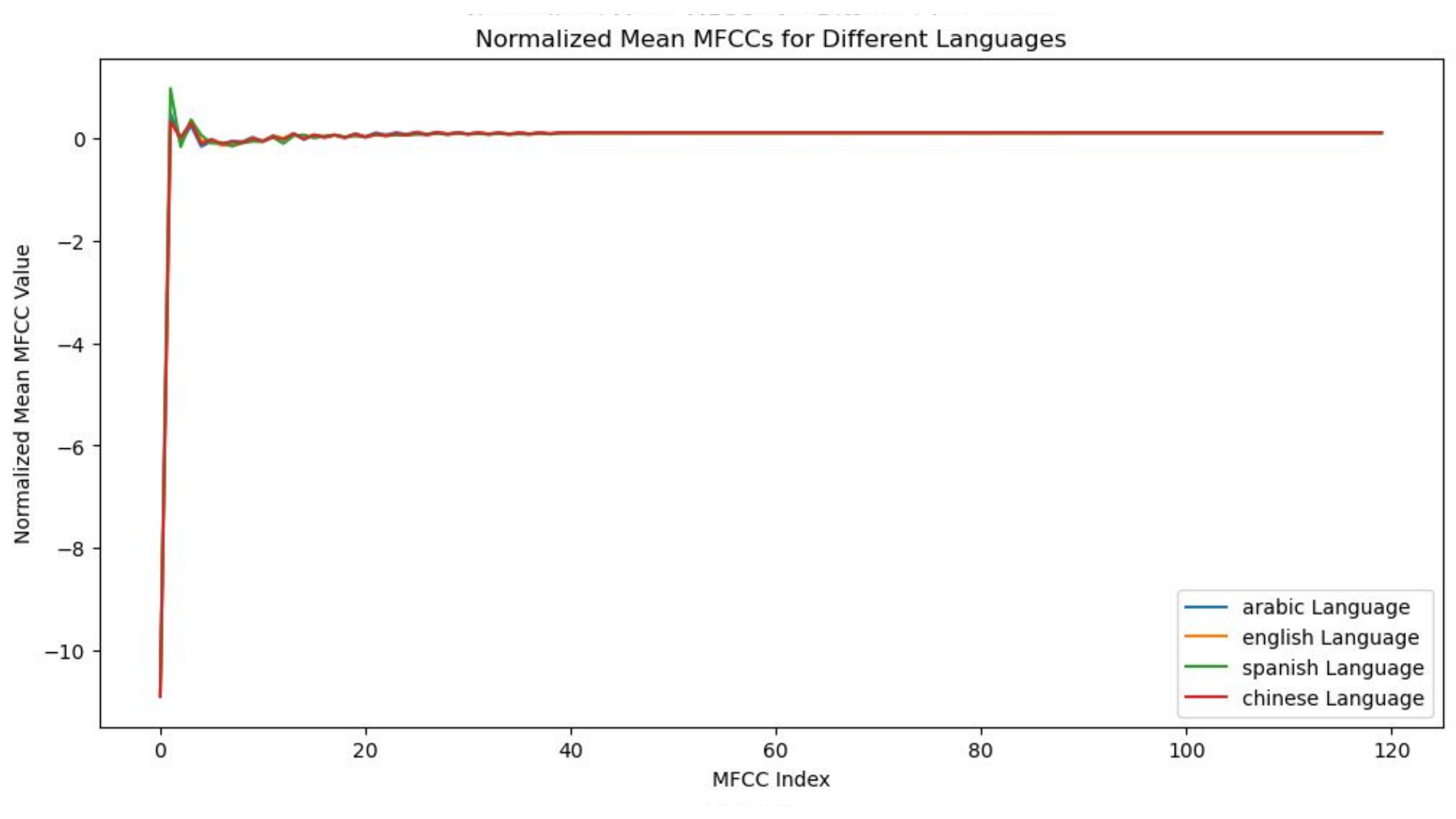
- Data Cleaning:
  - Checked for duplicates, null values
  - Made preliminary graphs of each of my columns

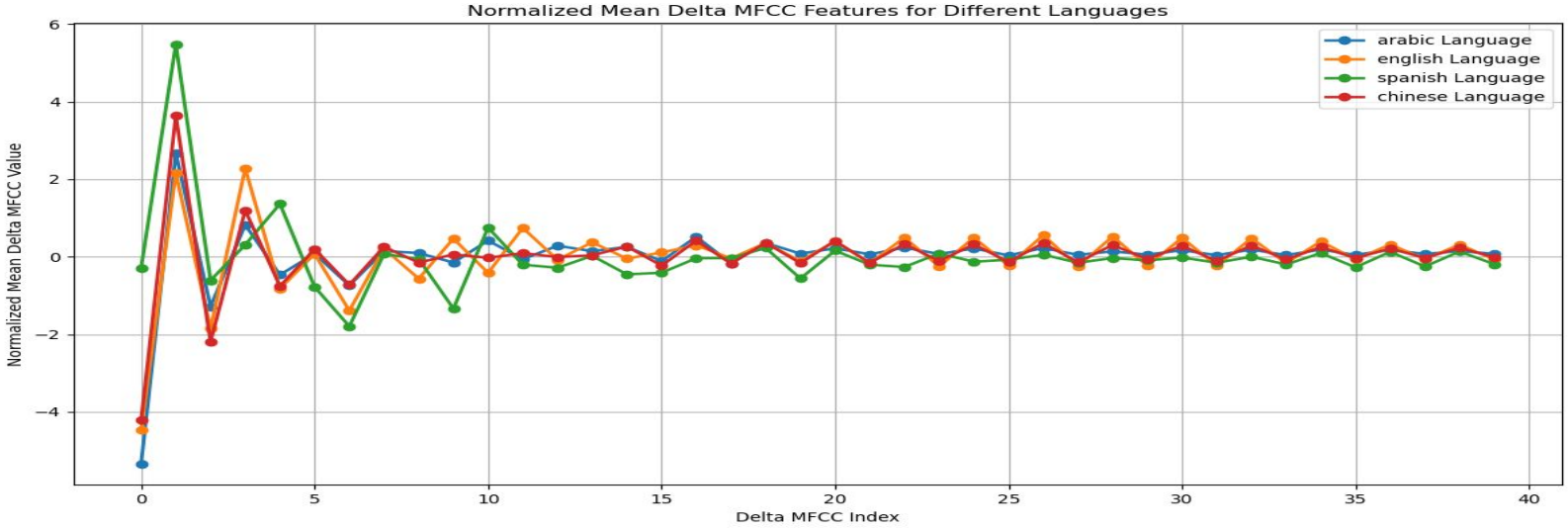
- Visualizations
  - Count of wav files per language
  - Audio wave graphs and Spectrum graphs
  - MFCC, Chroma, and Spectral Graphs

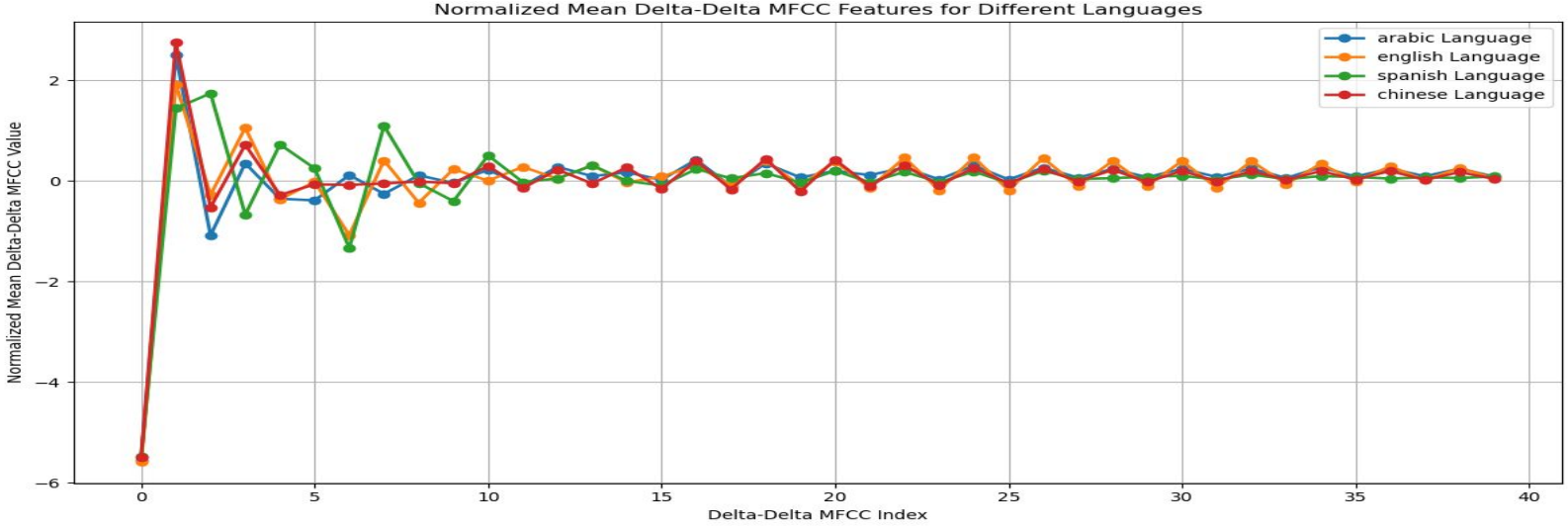

Count of Samples by Language
Number of Samples
70000
60000
50000
40000
30000
20000
10000
0
arabic
chinese
english
spanish
Language

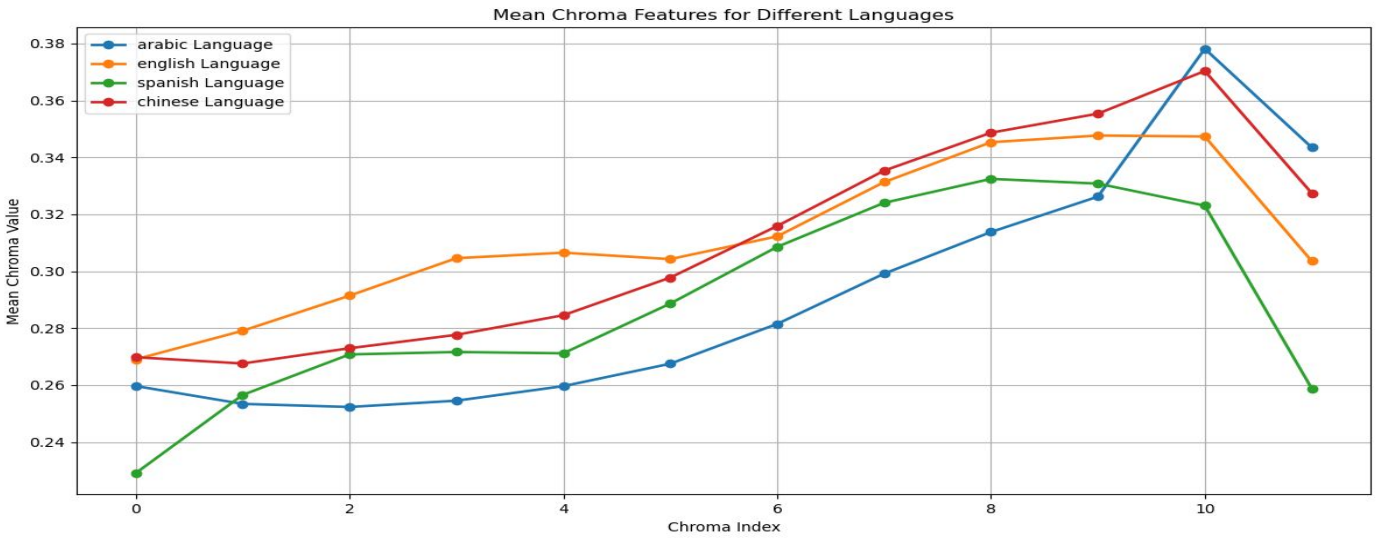

Waveform of common_voice_en_39586336.wav (Sample rate: 32000 Hz)
Amplitude
Time (s)
Waveform of common_voice_ar_19058307.wav (Sample rate: 48000 Hz)
Amplitude
Time (s)
Waveform of common_voice_zh-CN_18524189.wav (Sample rate: 48000 Hz)
Amplitude
Time (s)
Waveform of common_voice_es_34925862.wav (Sample rate: 32000 Hz)
Amplitude
Time (s)

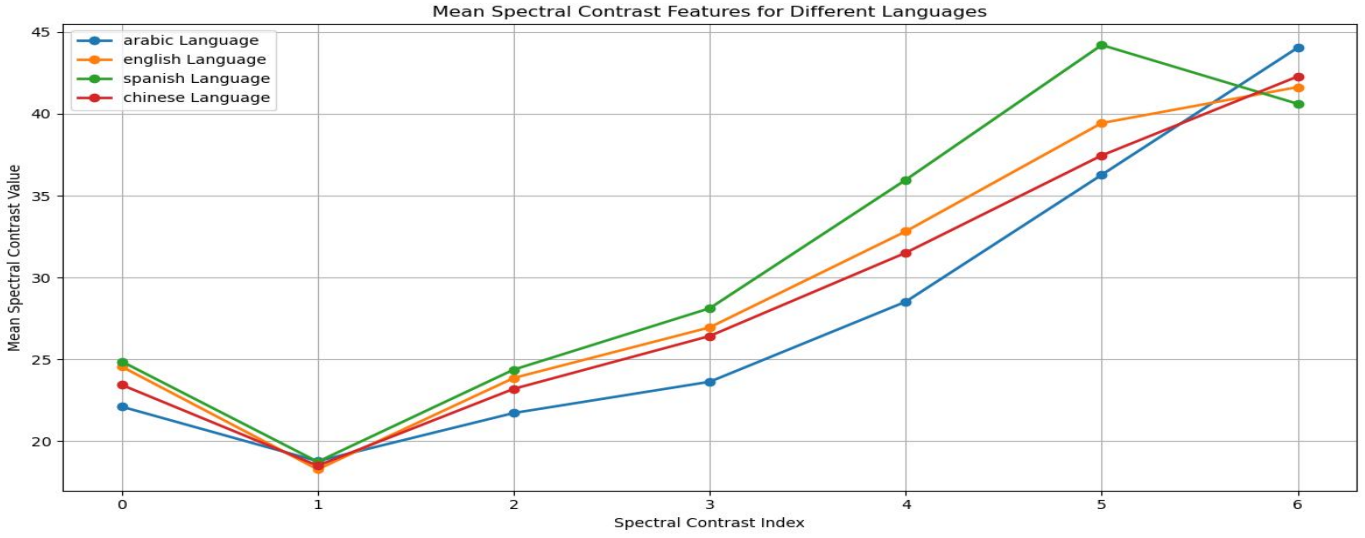

Mel-Spectrogram of common_voice_en_39586336.wav (Sample rate: 32000 Hz)
Frequency (Hz)
4096
2048
1024
512
0
0
0.5
1
1.5
2
2.5
3
3.5
4
Time (s)
+0 dB
-20 dB
-40 dB
-60 dB
-80 dB
Mel-Spectrogram of common_voice_ar_19058307.wav (Sample rate: 48000 Hz)
Frequency (Hz)
4096
2048
1024
512
0
0
0.5
1
1.5
2
2.5
3
3.5
4
Time (s)
+0 dB
-20 dB
-40 dB
-60 dB
-80 dB
Mel-Spectrogram of common_voice_zh-CN_18524189.wav (Sample rate: 48000 Hz)
Frequency (Hz)
4096
2048
1024
512
0
0
0.5
1
1.5
2
2.5
3
Time (s)
-20 dB
-40 dB
-60 dB
-80 dB
Mel-Spectrogram of common_voice_es_34925862.wav (Sample rate: 32000 Hz)
Frequency (Hz)
4096
2048
1024
512
0
0
0.5
1
1.5
2
2.5
3
Time (s)
+0 dB
-20 dB
-40 dB
-60 dB
-80 dB


Normalized Mean MFCCs for Different Languages
Normalized Mean MFCC Value
MFCC Index
0
−2
−4
−6
−8
−10
0
20
40
60
80
100
120
arabic Language
english Language
spanish Language
chinese Language


Normalized Mean Delta MFCC Features for Different Languages
arabic Language
english Language
spanish Language
chinese Language
Normalized Mean Delta MFCC Value
Delta MFCC Index


Normalized Mean Delta-Delta MFCC Features for Different Languages
arabic Language
english Language
spanish Language
chinese Language
Normalized Mean Delta-Delta MFCC Value
Delta-Delta MFCC Index


Mean Chroma Features for Different Languages
arabic Language
english Language
spanish Language
chinese Language
Mean Chroma Value
Chroma Index
0.38
0.36
0.34
0.32
0.30
0.28
0.26
0.24
0
2
4
6
8
10


Mean Spectral Contrast Features for Different Languages
arabic Language
english Language
spanish Language
chinese Language
Mean Spectral Contrast Value
Spectral Contrast Index
45
40
35
30
25
20
0
1
2
3
4
5
6

Next Steps in Data Processing and Model Development

- Further Data Processing
  - Check if there are other more effective methods of of capturing significant variance in the data using other methods (such as Zero Crossing Rate and Spectral Roll-off)
  - Explore Mel-Spectrograms more to see if they are more suitable to for my data.

- What potential Machine Learning models to use?
  - Start with simpler models like Random Forest and SVM to establish a baseline. If necessary, move on to more complex models like CNNs to improve accuracy.

The End!