The Language of Sounds

Spoken Language Identification

Ömer Faruk Satık
150210330

Nurullah Eren Acar
150220310

What presentation covers

1. PROBLEM DEFINITION
2. DATA COLLECTION
3. DATA PREPROCESSING
4. DATA AUGMENTATION
5. MODELS
6. CONCLUSION

Problem Definiton

Automatic spoken language identification is an important task needed in many fields, such as multilingual conversation systems or spoken language translation. With quality data, artificial intelligence models can perform this task successfully.

Data Collection

Youtube

- 4 languages
  - Arabic, Spanish, French, Turkish
- 109 different youtube videos
- Creative Commons license
- Nearly 6 hours of speech
- Download wav from youtube videos

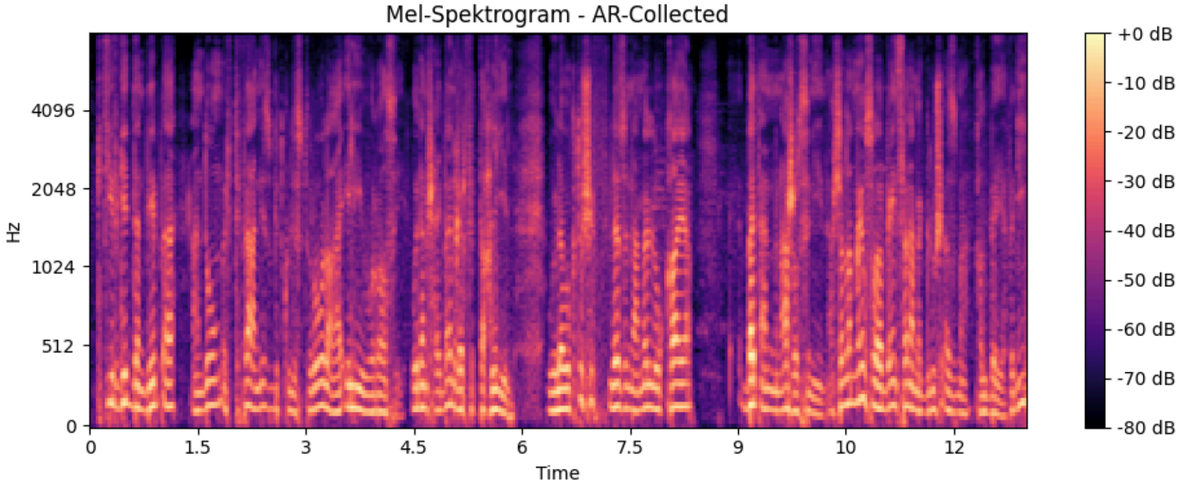
OpenSLR

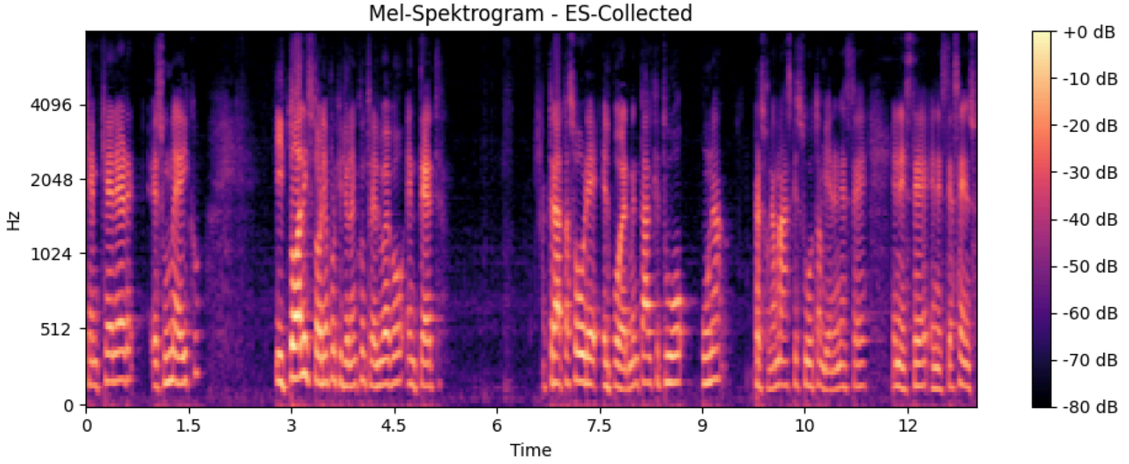
- 4 languages
- 40 hours of speech

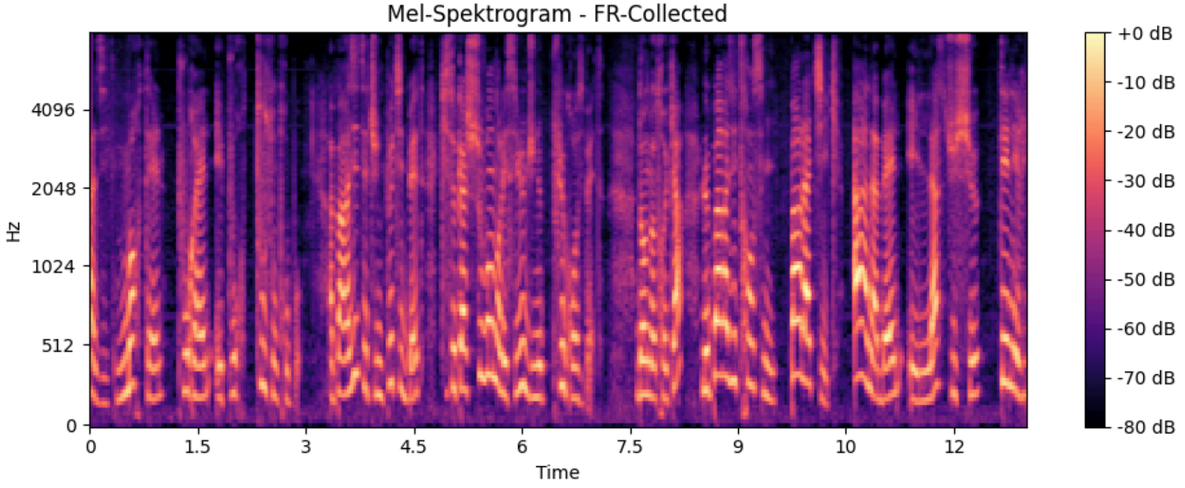
Data Preprocessing

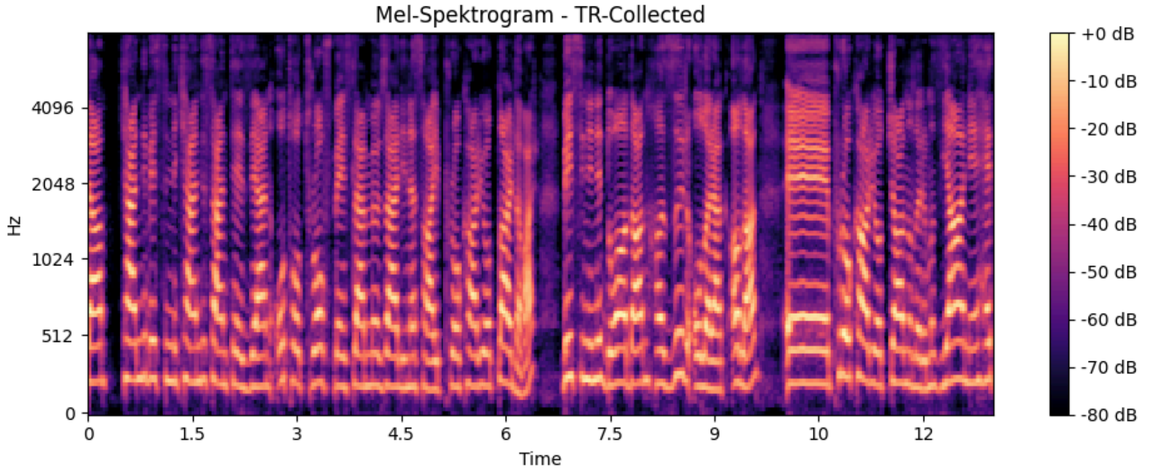
To ensure common data shape, each audio file was included in the Python environment using the librosa library with a **sample rate of 16000.**

- Clip only the talking parts
- Trim silent parts of audios
- Split data into 13-second segments
- Transform the audios from amplitude domain to frequency domain using Fourier transform (librosa.mel_spectrogram)


Mel-Spektrogram - AR-Collected
Hz
4096
2048
1024
512
0
0
1.5
3
4.5
6
7.5
9
10
12
Time
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB


Mel-Spektrogram - ES-Collected
Hz
4096
2048
1024
512
0
0
1.5
3
4.5
6
7.5
9
10
12
Time
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB


Mel-Spektrogram - FR-Collected
Hz
4096
2048
1024
512
0
0
1.5
3
4.5
6
7.5
9
10
12
Time
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB


Mel-Spektrogram - TR-Collected
Hz
4096
2048
1024
512
0
0
1.5
3
4.5
6
7.5
9
10
12
Time
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB

Data Augmentation

We aimed to increase the generalization ability of the models by augmenting the training data.

- **Pitch Shifting:** Changes the tone of the audio by increasing or decreasing its frequency
- **Adding Noise:** Adds a Gaussian noise to the audio
- **Time Masking:** Silences 1 random 0.2 second time period

CNN MODEL

Convolution Layers

```
conv2D = nn.Conv2d()
bn = nn.BatchNorm2d()
pool = nn.MaxPool2d(2)
spatial_drop = nn.Dropout2d(0.3)
```

```
conv2D = nn.Conv2d()
bn = nn.BatchNorm2d()
pool = nn.MaxPool2d(2)
spatial_drop = nn.Dropout2d(0.3)
```

```
conv2D = nn.Conv2d()
bn = nn.BatchNorm2d()
pool = nn.MaxPool2d(2)
spatial_drop = nn.Dropout2d(0.3)
```

```
conv2D = nn.Conv2d()
bn = nn.BatchNorm2d()
pool = nn.MaxPool2d(2)
spatial_drop = nn.Dropout2d(0.3)
```

Dense Layers

```
self.flatten = nn.Flatten()
self.fc1 = nn.Linear(25600, 256)
self.drop1 = nn.Dropout(0.4)
self.fc2 = nn.Linear(256, 64)
self.drop2 = nn.Dropout(0.4)
self.fc3 = nn.Linear(64, 4)
```

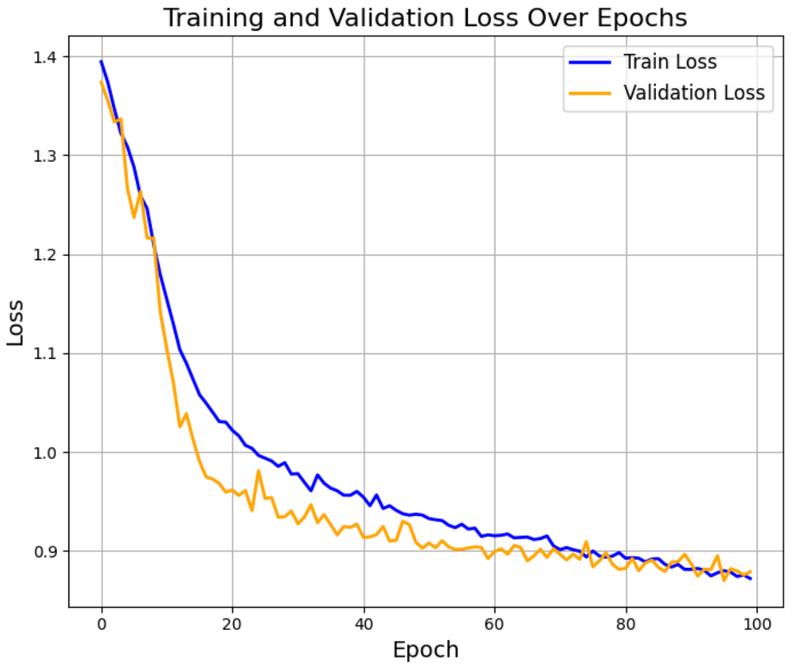
AR

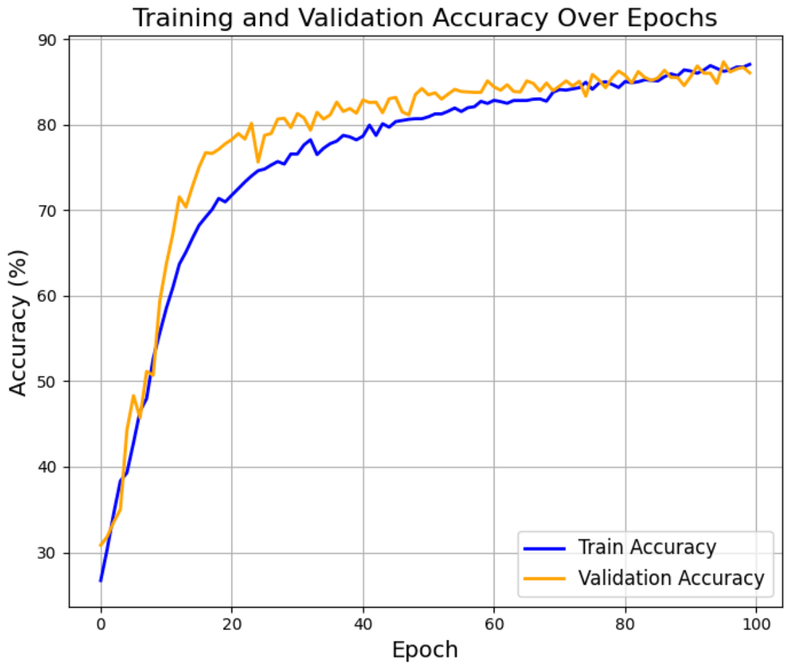
ES

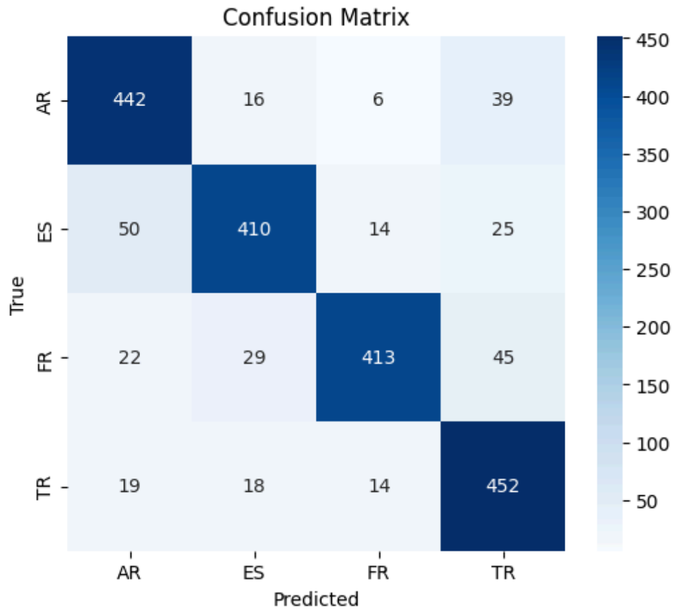
FR

TR

CNN MODEL





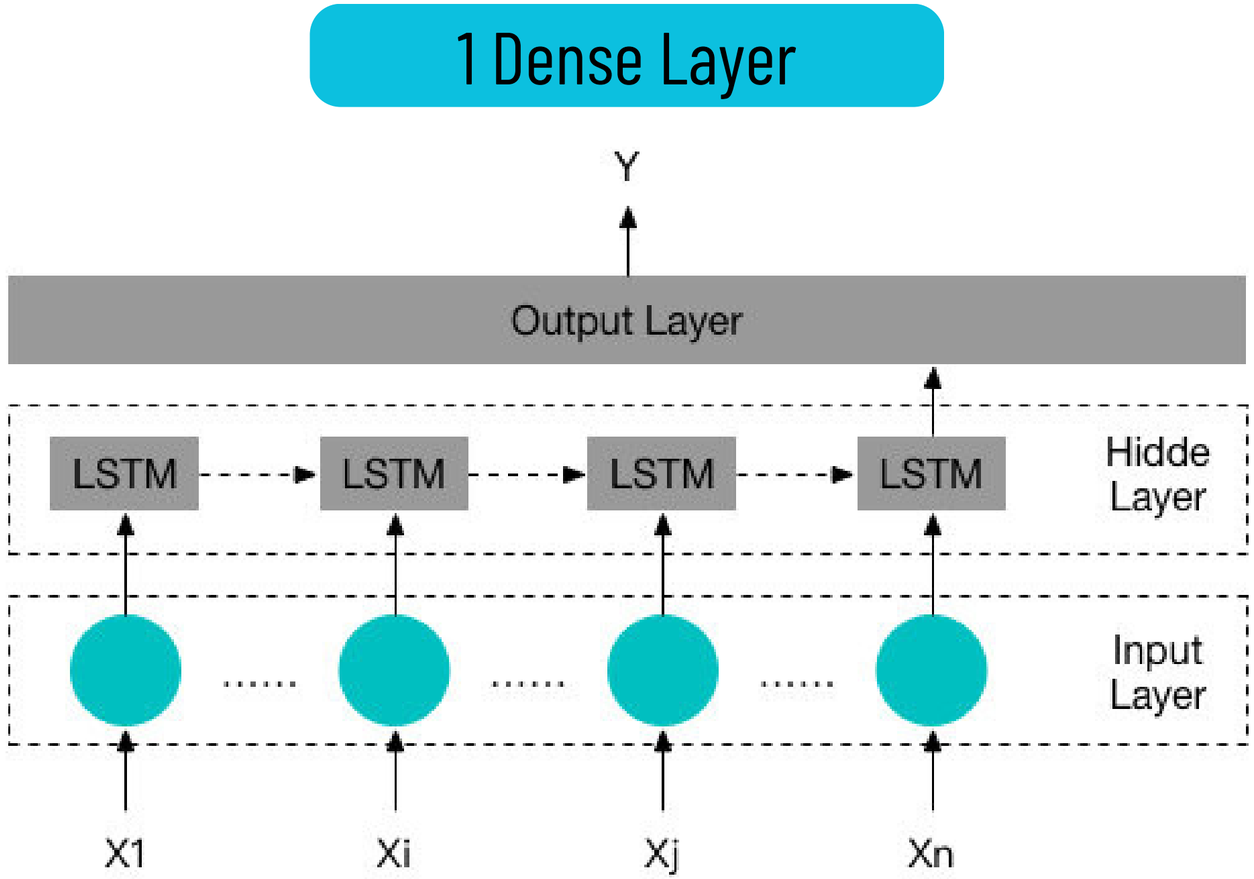


Test Accuracy

85%

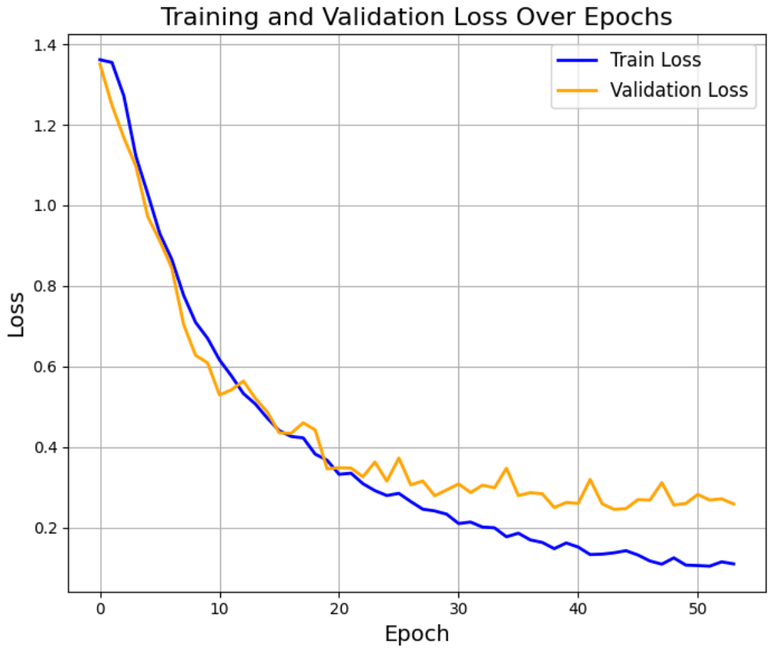
LSTM MODEL

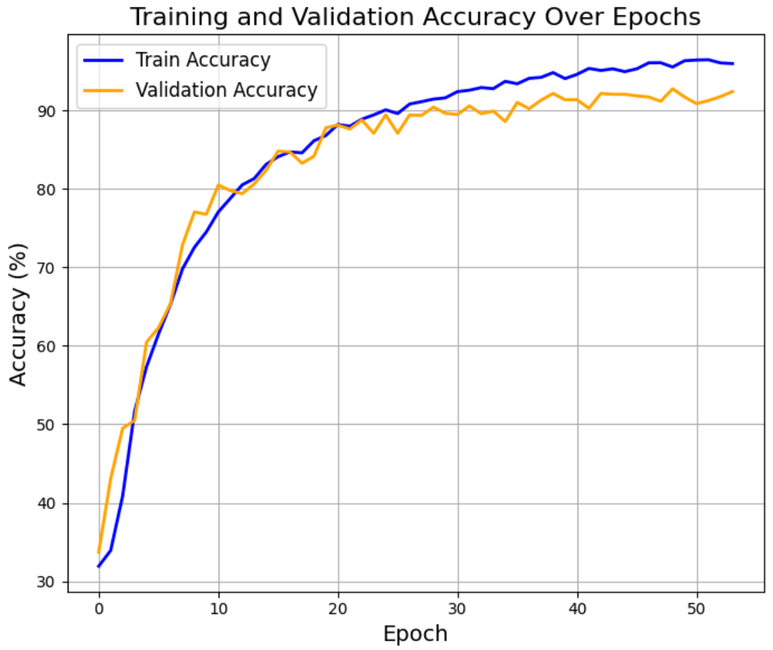
AR ES FR TR

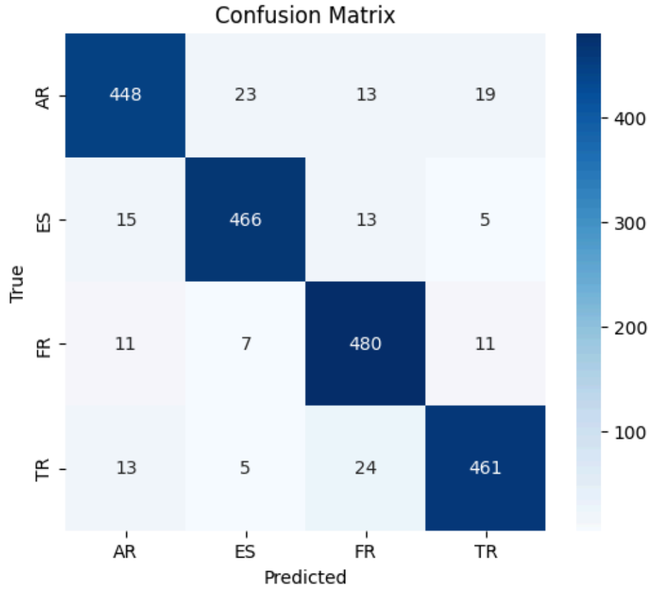


input_dim = 128
time_steps = 407
hidden_dim = 64

LSTM MODEL

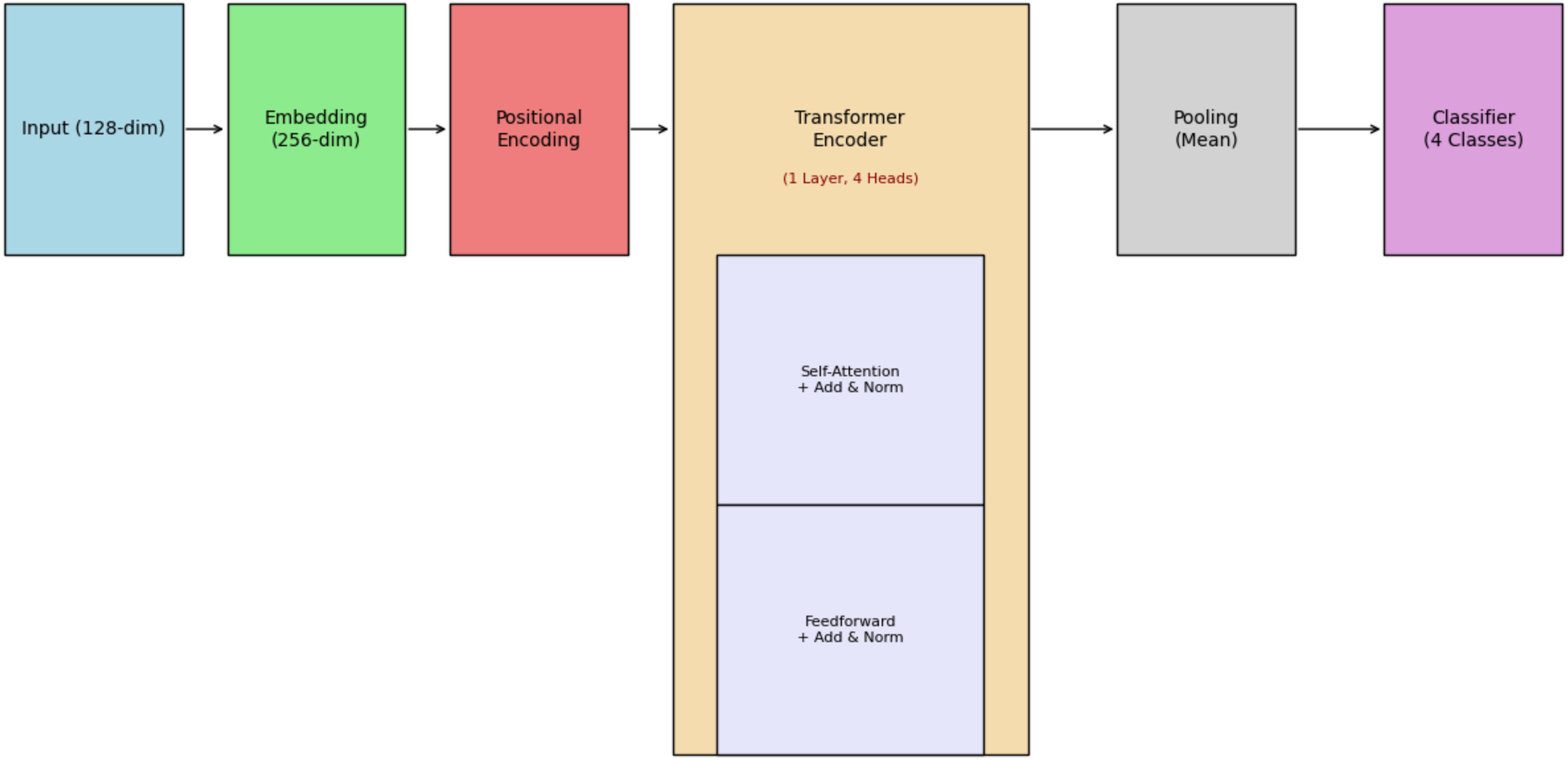




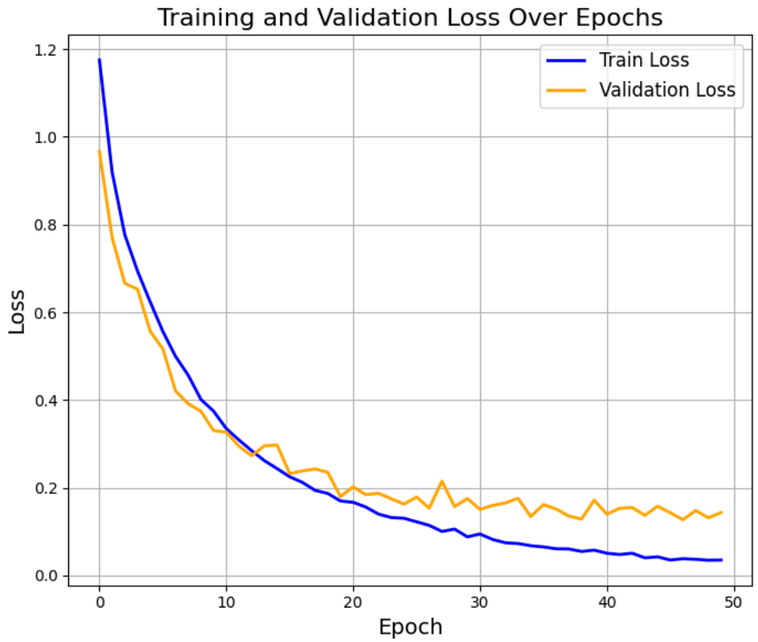


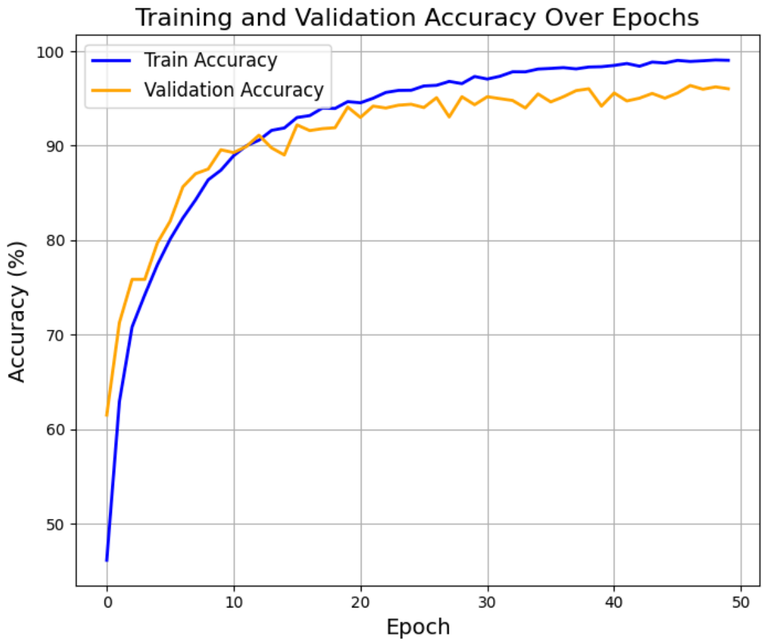
Test Accuracy
92%

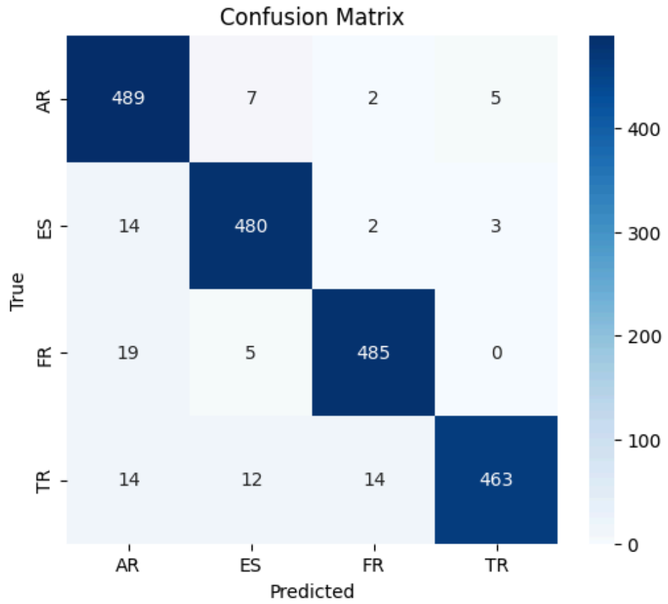
Transformer MODEL



Transformer MODEL







Test Accuracy
95%

Conclusion

- Mel spectrograms can show the difference between spoken languages quite well
- Since **LSTM** and **Transformer** models treat data as time series, they can better perceive the time-dependent features of different languages.
- **Transformer** model shows the best performance
- This study is promising for identifying across more languages