Corsound Interview Practical Test

Mor Zahavi

Corsound AI

Table of contents

- The goal is to fit a classifier to distinguish real speech from fake.
- A solution is expected to be a DL model that having an audio input (waveform) outputs a score with associated threshold to classify real/fake speech.
- As a target metric we suggest using EER (equal error rate)
- Dataset ASVspoof 2019

EER Definition

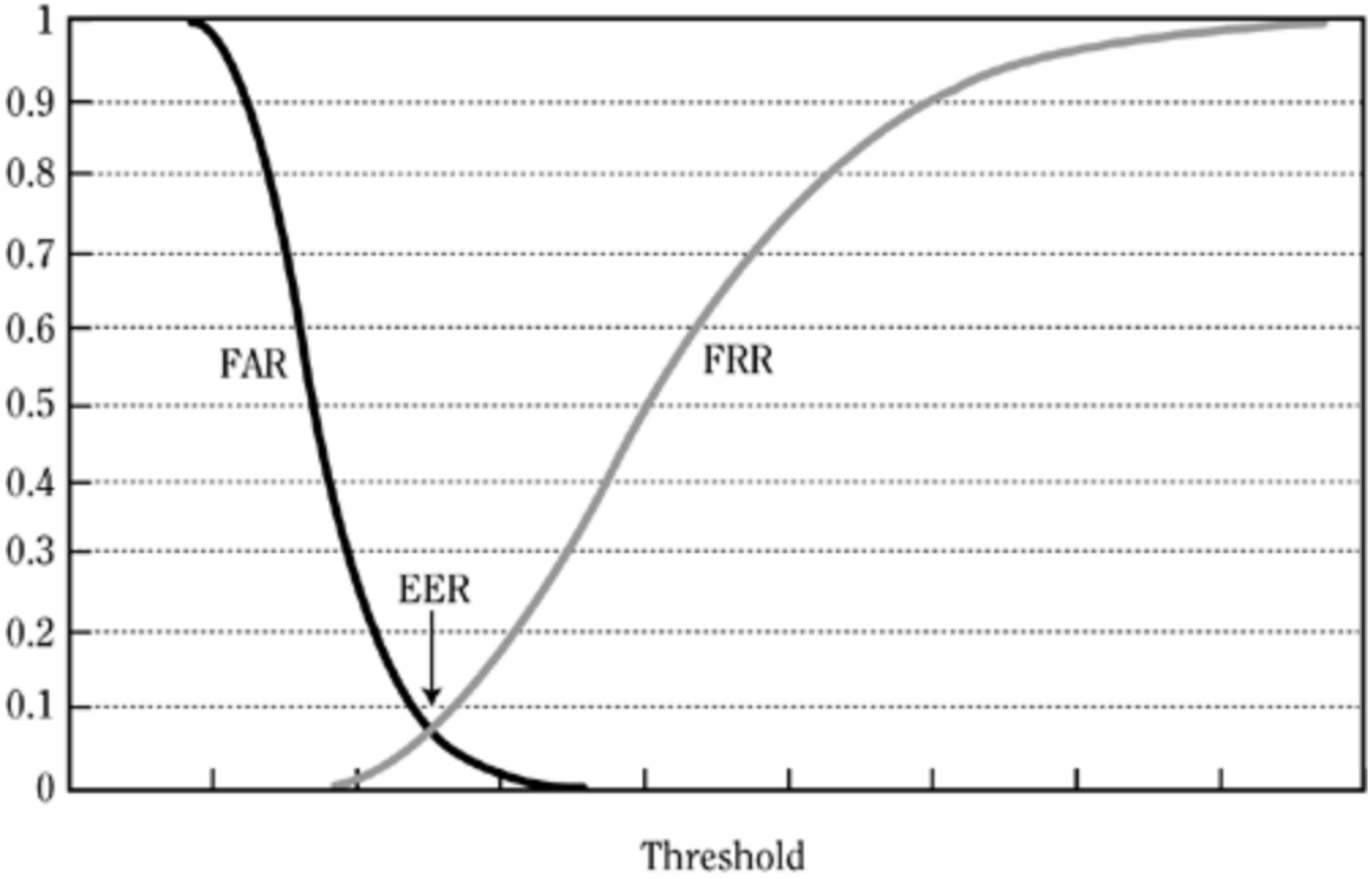
- EER is an important metric since it ensures that the system is able to accurately identify and verify individuals
- A low EER can lead to false rejections or acceptances

The Equal Error Rate (EER) is defined as:

$$\text{EER} = \arg \min_{\text{threshold}} |\text{FAR}(\text{threshold}) - \text{FRR}(\text{threshold})|$$

Where:
- FAR(threshold) represents the False Accept Rate at a given threshold.
- FRR(threshold) represents the False Reject Rate at a given threshold.
- arg min finds the threshold that minimizes the absolute difference between FAR and FRR.

EER visualisation



FAR, FRR and EER [1]

Sources

I relied heavily on

- https://www.kaggle.com/code/awsaf49/asvspoof-2019-tfrecord-data
- https://www.kaggle.com/code/awsaf49/fake-speech-detection-conformer-tf

Meta Data

I converted the files to tfrecords

- speaker_id : LA_**, a 4-digit speaker ID
- filename : LA_**, name of the audio file
- system_id : ID of the speech spoofing system (A01 - A19), or, for real speech SYSTEM-ID is left blank ('-')
- class_name : bonafide for genuine speech, or, spoof for fake/spoof speech
- target : 1 for fakespoof and 0 for real/genuine

Audio:

- Random Noise
- Random TimeShift
- Random CropOrPad
- Audio Trim

Spectogram:

- Random TimeMask
- Random FreqMask
- CutMix
- MixUp

Conformer Model

- The conformer model uses CNN and transformers to combine harvesting global and local features
  - Feed-forward module
  - Self-attention module
  - Convolution module
  - Layer normalization module
- The backbone used is ImageNet

Confusion Matrix

```
>> Confusoin Matrix
```