Deep Learning Project

KYRIAKI BEI

The Project

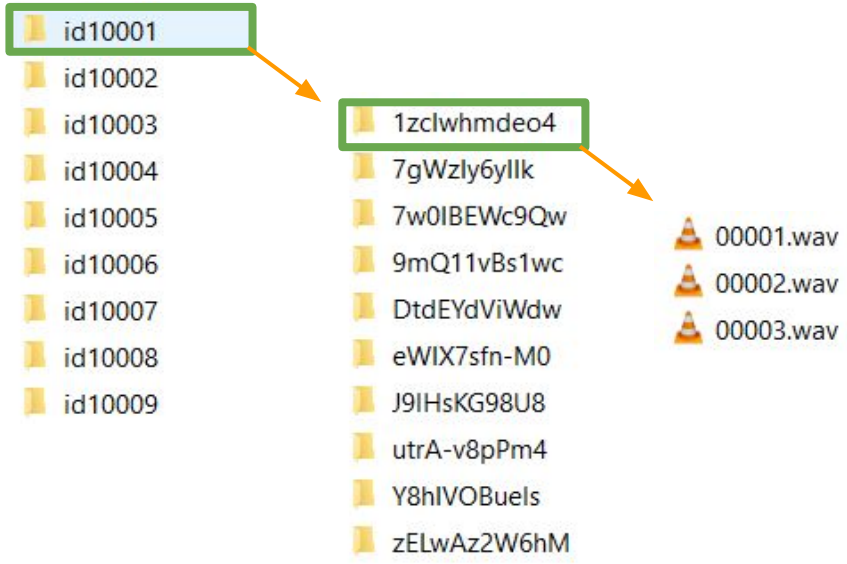
Subject

Speaker Identification by Voice.

Dataset

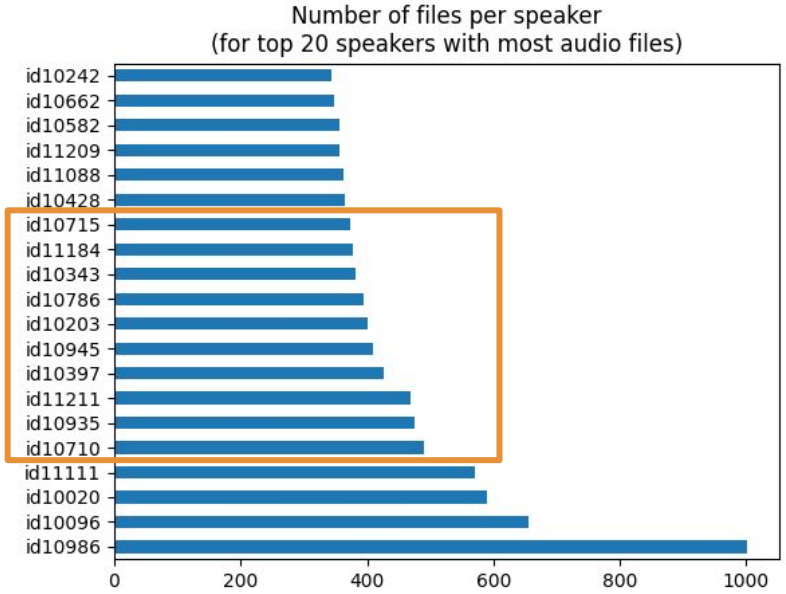
VoxCeleb1: An audio dataset consisting of over 100,000 short clips of human speech for 1.251 celebrities, extracted from interview videos uploaded to YouTube.

	dev files
# speakers	1,211
# audio files	148,642



File Handling

From the **1,211** total speakers, **10** were selected for the classification task according to their number of audio files.



Dictionary that maps the 10 speakers' ids to their audio files


```
{
    "id10710": [
        "data\\voxceleb_data\\wav\\id10710\\230QPHWW7fM\\00001.wav",
        "data\\voxceleb_data\\wav\\id10710\\230QPHWW7fM\\00002.wav",
        "data\\voxceleb_data\\wav\\id10710\\230QPHWW7fM\\00003.wav",
        "data\\voxceleb_data\\wav\\id10710\\230QPHWW7fM\\00004.wav",
        "data\\voxceleb_data\\wav\\id10710\\230QPHWW7fM\\00005.wav",
        "data\\voxceleb_data\\wav\\id10935\\Z1x9xyh8NjA\\00009.wav"
    ],
    "id11211": [
        "data\\voxceleb_data\\wav\\id11211\\0LUUizW5U48\\00001.wav",
        "data\\voxceleb_data\\wav\\id11211\\0LUUizW5U48\\00002.wav",
        "data\\voxceleb_data\\wav\\id11211\\6mnL46vX6NY\\00001.wav",
        "data\\voxceleb_data\\wav\\id11211\\6mnL46vX6NY\\00002.wav",
```

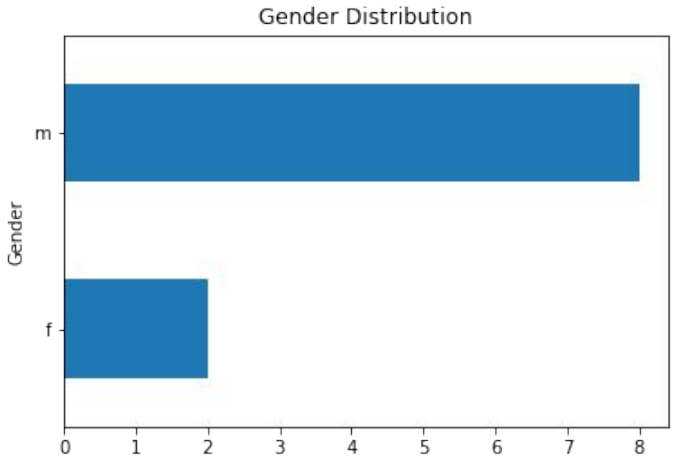
The Dataset

The dataset used for classification consists of **10 speakers** and **4,195** audio files in total.

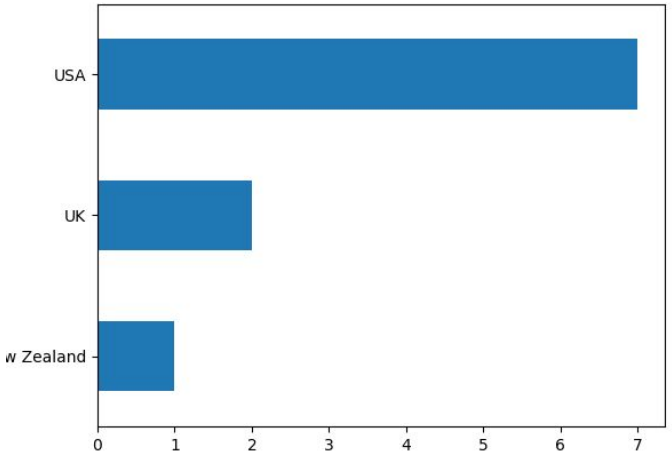
Selected speakers

1. David Attenborough
2. Gloria Steinem
3. J.J. Abrams
4. Louis C.K.
5. Lucie Arnaz
6. Meat Loaf
7. Peter Jackson
8. Quentin Tarantino
9. Tom Hooper
10. Vince Gilligan

Genders Distribution

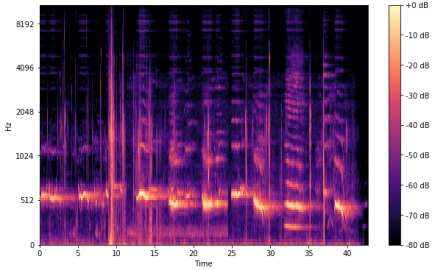


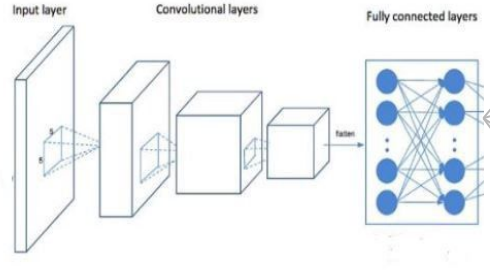
Nationalities Distribution

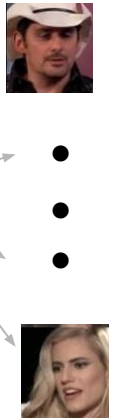


Workflow

STFT







waveform

- same sample rate
- same # channels

mel spectrogram

CNN

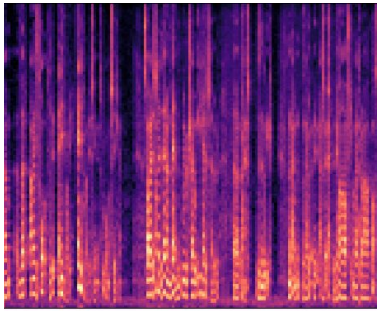
FC layer

Audio Preprocessing

sample rate = 16000 kHz

mono signal

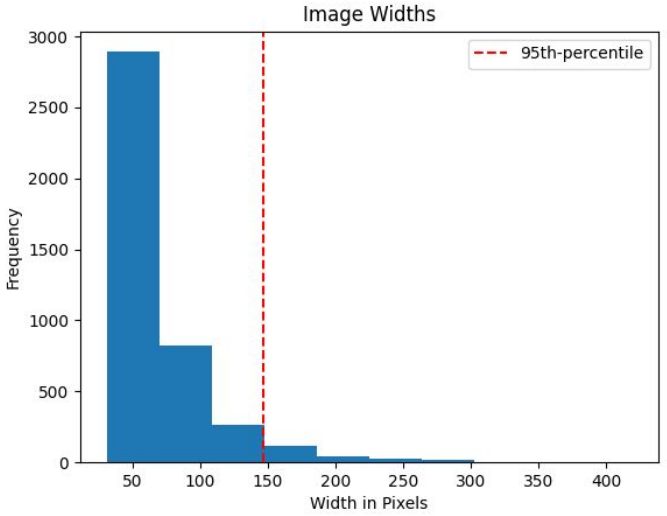
mel spectrogram



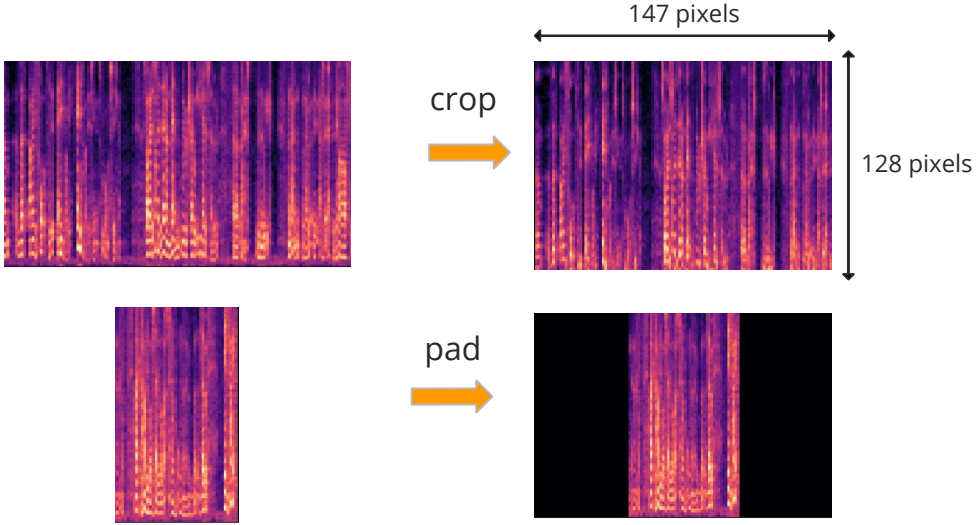
Audio Preprocessing

- optimal image width for spectrogram images found

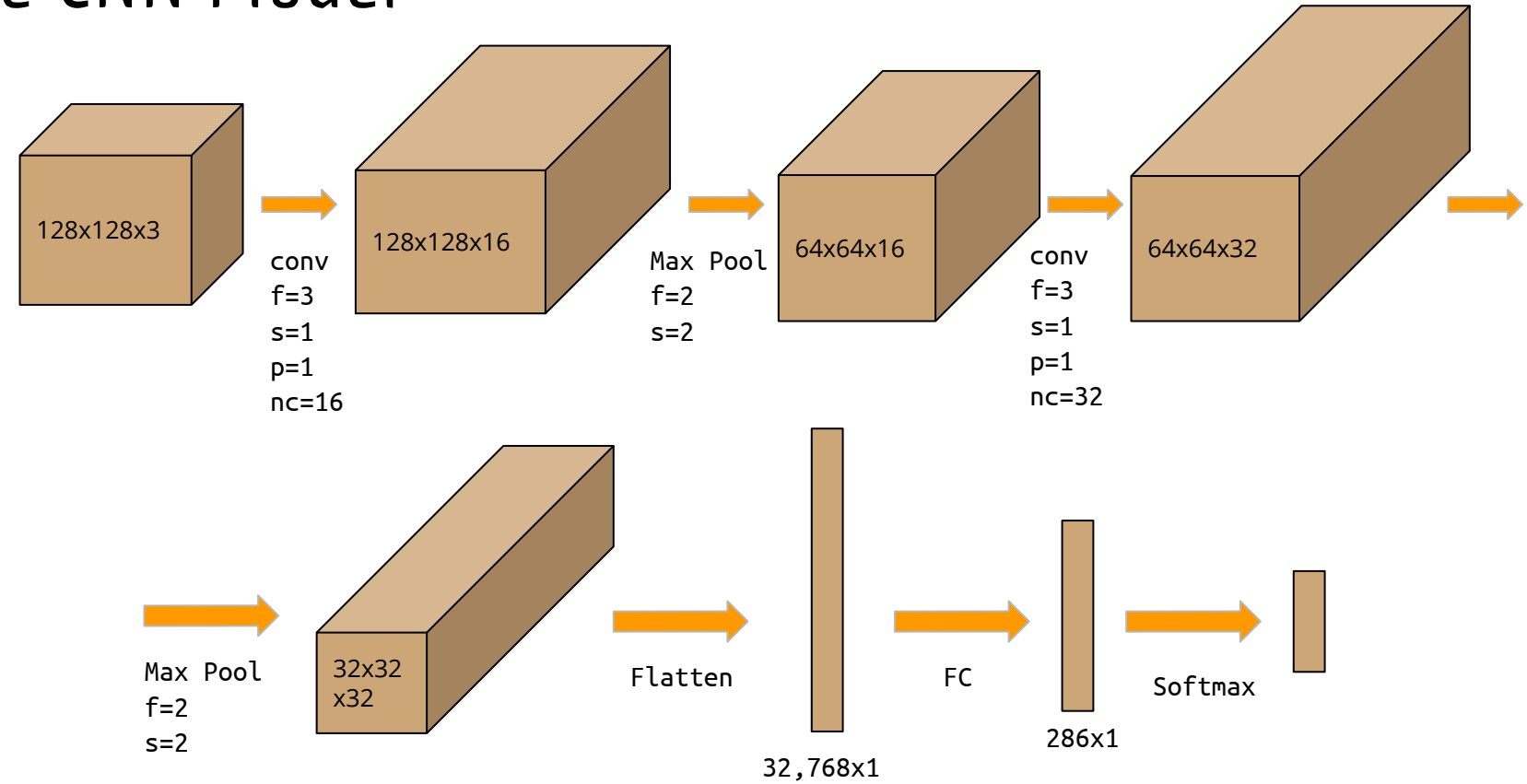
  optimal width = 95th-percentile of image widths



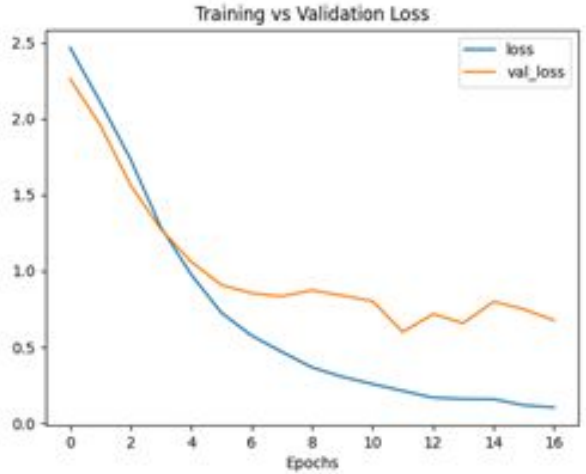
- crop or pad images to have width equal to the optimal image width

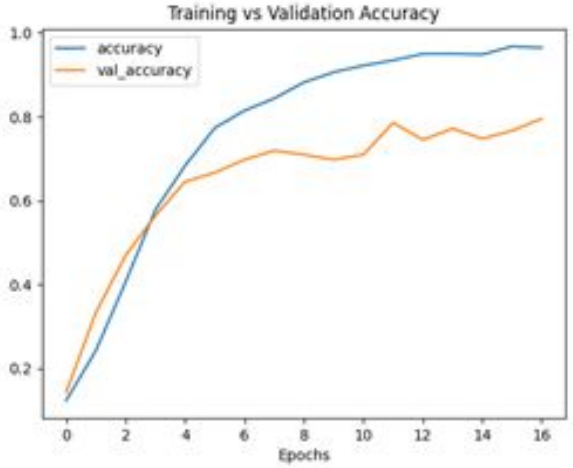


The CNN Model



Results





Train loss: 0.105
Train accuracy: 96.4%

Val loss: 0.675
Val accuracy: 79.5%

Test loss: 0.389
Test accuracy: 86.4 %

Some Predictions

- Predicted speaker **Gloria Steinem** with probability **99.8 %.** (True speaker: Gloria Steinem)
- Predicted speaker **Tom Hooper** with probability **99.0 %.** (True speaker: Tom Hooper)
- Predicted speaker **Quentin Tarantino** with probability **83.8 %**. (True speaker: Quentin Tarantino)