Capstone Proposal

Domain Background

The project belongs to the audio speech emotion recognition domain.
In particular, from a set of audio files, will be extracted the main features and with Machine Learning techniques will be tempted to classify those by emotional categories.

Problem statement

Emotion recognition algorithm is mainly used to:

- Helping children with autism
- Helping robot to better interact with people
- Monitoring signs for system security
- Marketing analysis

Datasets and inputs

The database used in the project has been selected on Kaggle based on the quality, quantity of data. The previously developed projects can be also considered a benchmark with which compare the results of the final model.

The dataset can be retrieved to the following link:
https://www.kaggle.com/uwrfkaggler/ravdess-emotional-speech-audio/kernels

Future of the Dataset:

- 1440 files
- 24 actors (12 female, 12 male- neutral North American accent)
- 60 trials per actor
- Speech emotions categories:
  - calm,
  - happy,
  - sad,
  - angry
  - fearful
  - surprise
  - disgust

The dataset contains 192 instances for each type of the emotion divided in 2 type of intensity, 96 expression normal and 96 strong. In additional are included other 96 neutral expression instances.

The final dataset results really balanced.

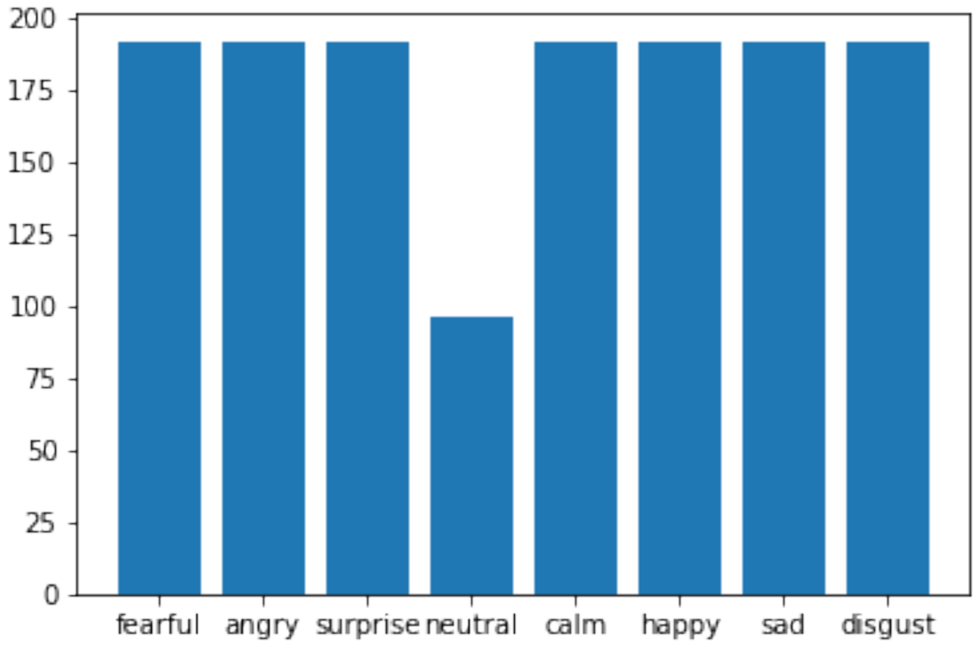


Figure 1

Solution statement

Goal of the project is to rightly classify the emotional intention of a speech as most as possible. In order to achieve this will be tested two different models.

1. CNN model with a convolutional layer and a pool layer described in following paper form Stanford University:
   Balakrishnan, Anusha, and Alisha Rege.
   Reading Emotions from Speech Using Deep Neural Networks.
   Stanford, 2017, pp. 1–8, *Reading Emotions from Speech Using Deep Neural Networks*.
   The model will be developed with Pytorch.

2. CNN model composed with 4 convolutional layer developed proposed in the following Kaggle notebook but tested on another dataset:
   https://www.kaggle.com/ejlok1/audio-emotion-part-6-2d-cnn-66-accuracy
   This model is developed with Tensorflow.

Goal of the project is to use similar features used for these experiments, test the models with a bigger and balanced dataset to understand if there is space of improvement.

Benchmark model

Because of the nature of the data is reasonable to compare the models tested with a linear logistic regression.

Evaluation metrics

Because the balance level of the dataset the metric used to evaluate the model outcome will be the accuracy.
These parameters will be also compared with the original research and notebook to understand if the models are generally reliable and applicable.

Project design

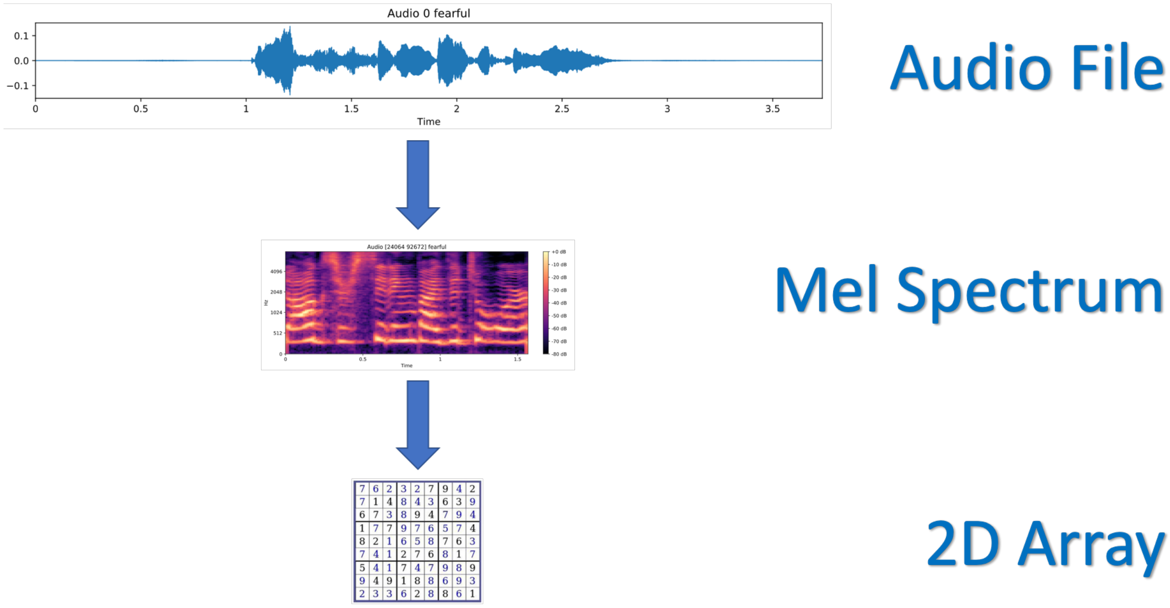
In order to select meaningful data based on which predict the emotions will be first selected the main feature used in the two considered studies: the MFCC spectrum.

In order to obtain this will be necessary to:

1. Take the Fourier transform of (a windowed excerpt of) a signal.
2. Map the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows.
3. Take the logs of the powers at each of the mel frequencies.
4. Take the discrete cosine transform of the list of mel log powers, as if it were a signal.
5. The MFCCs are the amplitudes of the resulting spectrum.[1]

To retrieve the MFCC will be used "librosa" a Python library able to get an audio file and return an array of set of frequencies sampled by time frame.
In detail for each audio file will be returned an array [*n x m*] where *n* is the number of the frequencies considered in each temporal window and *m* the number of time sampling.

[1]
Wikipedia 2020
Mel-frequency Cepstrum
https://en.wikipedia.org/wiki/Mel-frequency_cepstrum



7	6	2	3	2	7	9	4	2
7	1	4	8	4	3	6	3	9
6	7	3	8	9	4	7	9	4
1	7	7	9	7	6	5	7	4
8	2	1	6	5	8	7	6	3
7	4	1	2	7	6	8	1	7
5	4	1	7	4	7	9	8	9
9	4	9	1	8	8	6	9	3
2	3	3	6	2	8	8	6	1

Figure 2

As starting will be set the following parameters:

- n = 40 band frequencies
- m = 345
- sampling rate of 44100

From each audio file will be considered and extracted the Mel spectrum features, will be considered 40 bands for 4 second of sampling composing an input matrix of 40 x 345 for each audio file.

Two notebooks will be developed, in the first case the model will be developed from sketch based on the paper guidelines, in the second case the model will be introduced and adapted to the new dataset.
The first model (Fig3) is composed by 1 Convolutional layer, 1 Pool layer and 8 fully connected layer for the classification, 1 for each emotion.
The second model (Fig4) is composed by 4 convolutional layers with batch normalization and pool for each passage.

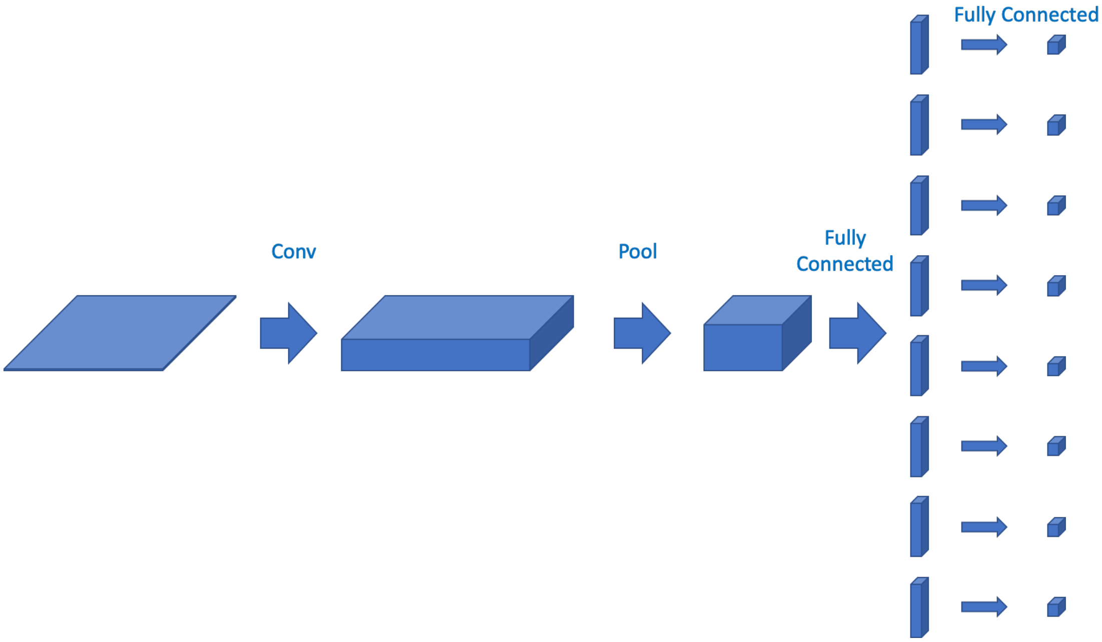


Figure 3

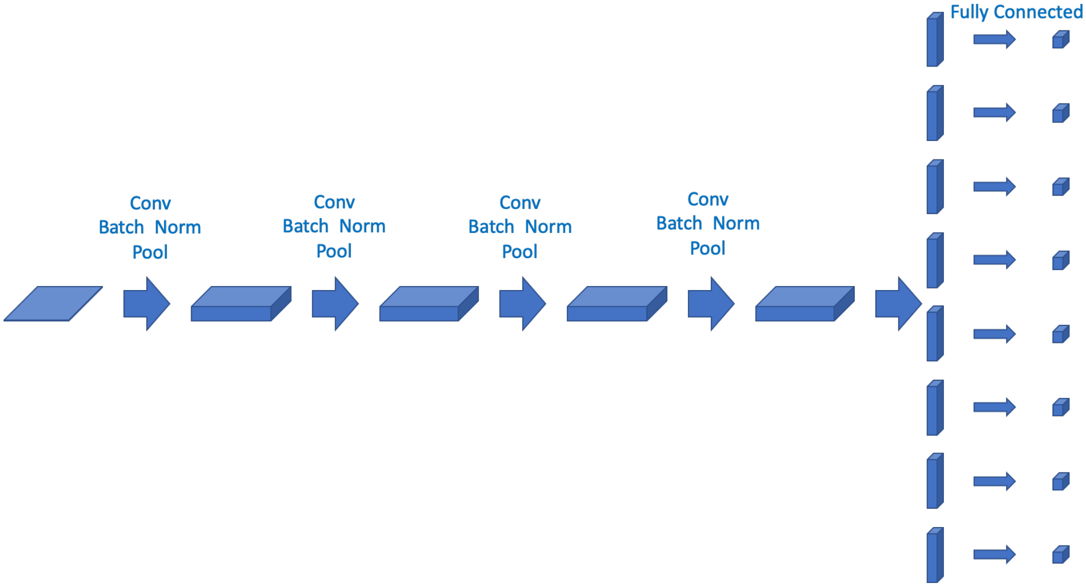


Figure 4

Finally, different hyperparameters will be apply in order to understand if there is space of improvement. In the end the different metrics will be compared.