Project Report

Problem Background

Project Overview

The project that is going to be developed has the goal to recognize the prevalent emotion in a given audio file containing a speech.

The problem is going to be tackled using the combination of the Digital Signal Process (DSP) algorithm to extract some potential features from the audio files. Finally, because of the features of the source involve in the time dimension, it is appropriate to use a Convolutional Neural Network as a Machine Learning technique to attempt to extract significant patterns and make predictions.

Problem Statement

The idea at the base is to split the input audio file into little slices and for each slice extracting the sounds components and their intensity in a way similar like the ears do. From that, is possible to analyze the trend of these components and try to identify some pattern related to the emotional information.

The algorithm should be able to classify the main emotion expressed in the speech.

Metrics

To define the efficacy of the model, considering the features of the initial dataset, will be evaluated total accuracy of the predictions.

The Set Up

Data Exploration

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 to 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 emotion divided into 2 types of intensity, 96 expressions normal and 96 strong. In additional is included other 96 neutral expression instances.

The final dataset results really balanced.

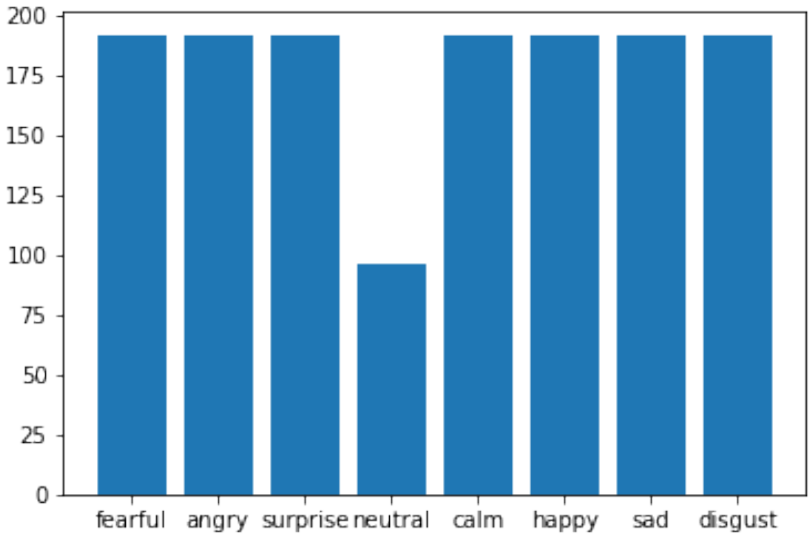
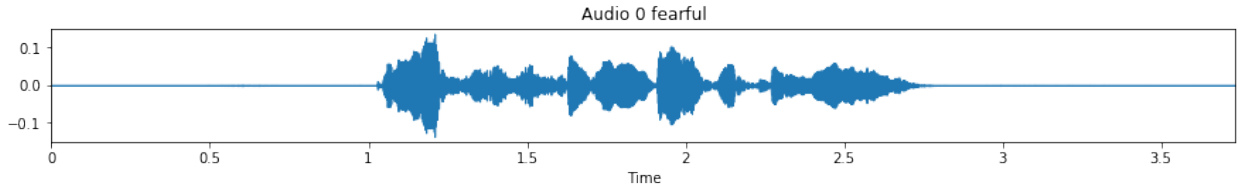


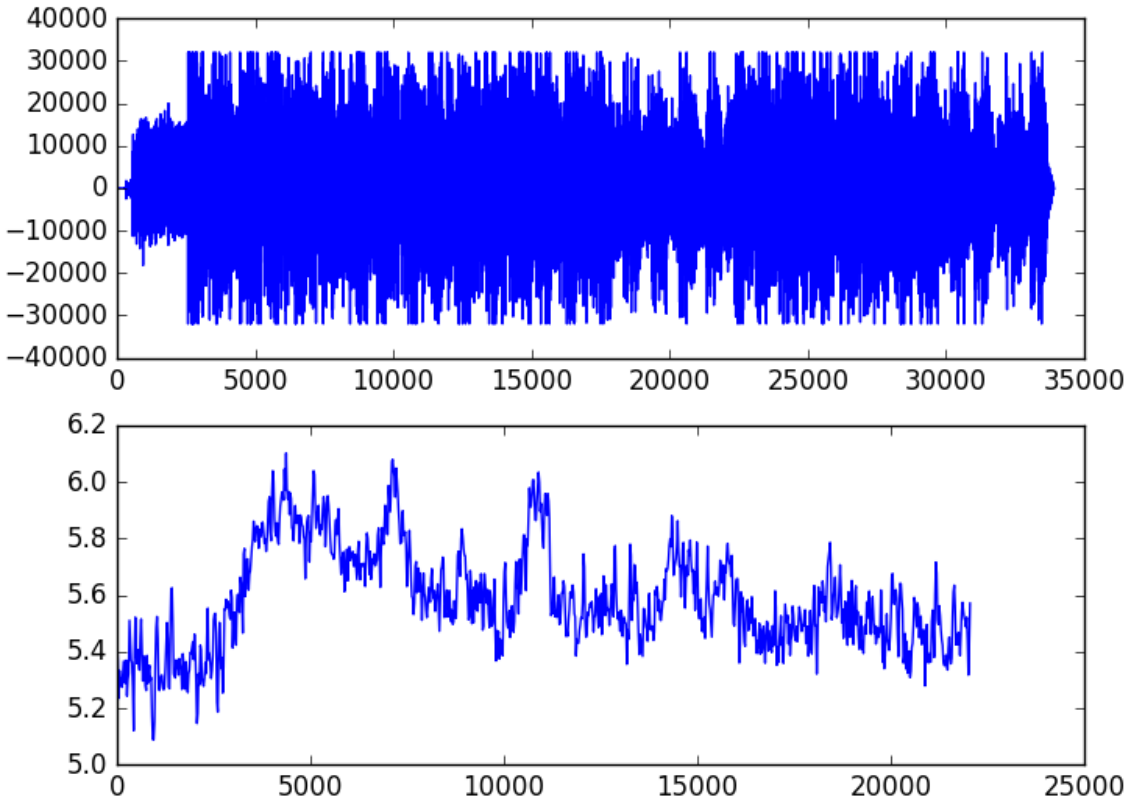
Figure 1

Exploratory Visualization

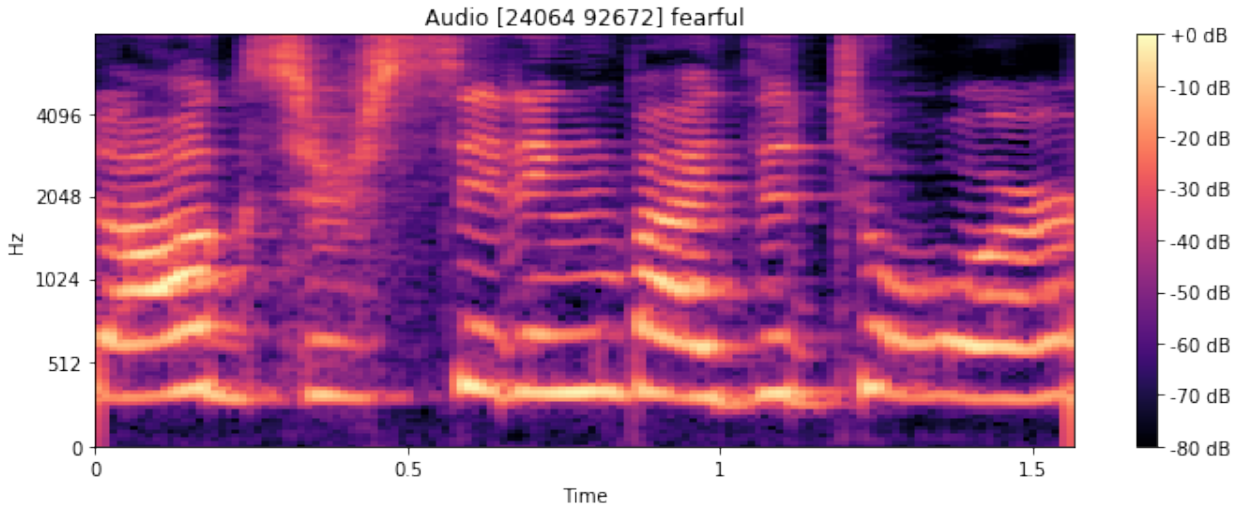
Being audio files, the digital representation of the source information is a sequence of 16 bits numbers that describ audio sound wave:



As described above the source will be split in different slices and each break up into the simple sounds that compose it (frequencies) and their intensity obtain the spectrum of the frame:



Each spectrum will be then transformed and equalized multiplying its components by a function that simulates the way that ears perceive the sounds obtain the better-known MEL spectrum. Combining all the slice will result in the following Mel Spectrogram:



The Mel Spectrogram represent the input data for the Machine Learning Model.

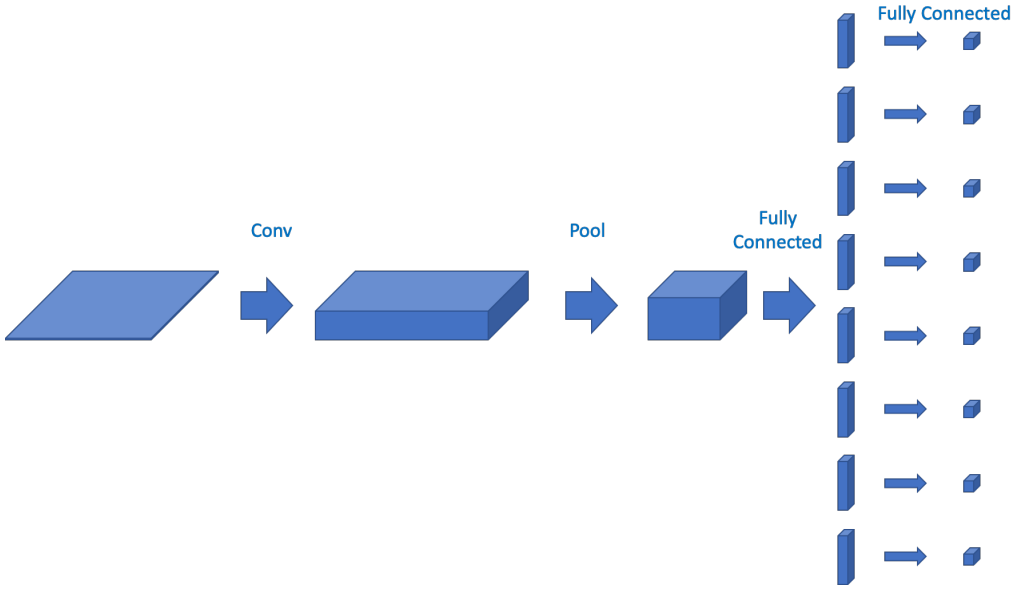
Algorithms and Techniques

Based on the researches done, two models have been considered to develop.

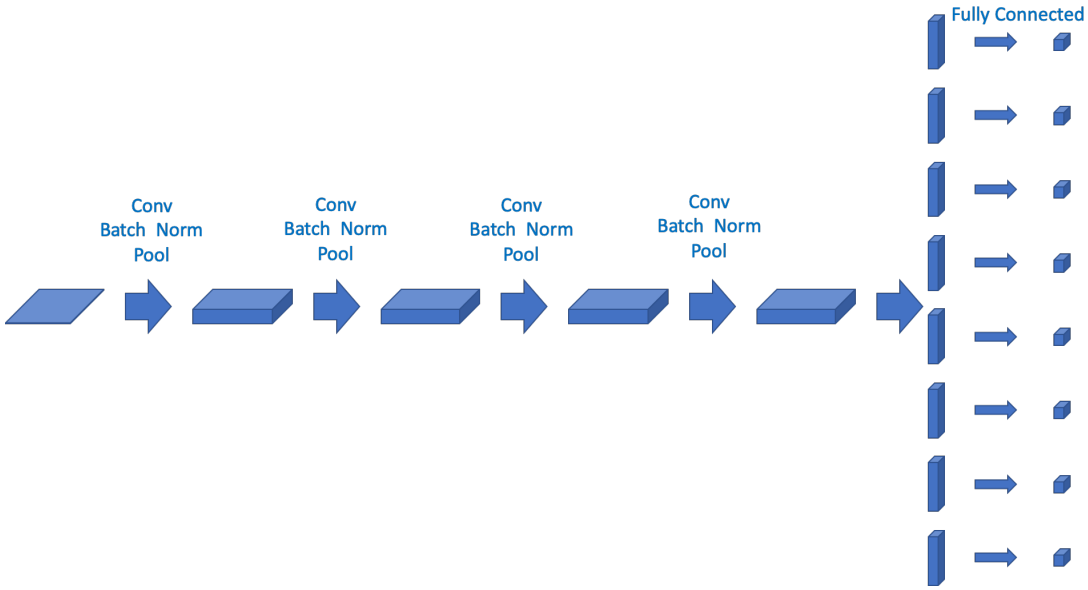
The first model is contained in the following paper:

> 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 is composed of 1 2D Convolutional layer, 1 Pool layer and 8 fully connected layers for the classification, 1 for each emotion. This model will be developed in Pytorch.



The second model proposed in the following Kaggle notebook promising good results https://www.kaggle.com/ejlok1/audio-emotion-part-6-2d-cnn-66-accuracy is composed of 4 convolutional layers with batch normalization and pool for each step. This will be developed in Keras.



Benchmark

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

Implementation

Data preprocessing

The preprocessing phase is common to both models.
Because the features are directly extracted by the audio files and the dataset is well balanced and already guaranteed in quality by Kaggle, no anomalies were found, and no cleaning process was required.
The dataset contained the references to the audio files and their names are loaded and split in 75% for the training set and 25% for the test. The training set is itself divided into training for 75% and the other 25% for the validation.
Then, for each is extracted the emotion label from the name and after wrapped it in a 4 seconds file container the Mel spectrogram is computed with the 'librosa' library.
After that, the values are normalized and ready for the training process.

Data Parameters

In the process two main variables required to be set to compute the MEL spectrum, these are the audio sample frequency and the number of the frequency bands calculated for each spectrum of the audio. The audio sample frequency is set to 44100, a standard value that guarantees to don't lose quality and information. According to the consulted literature, the number of bands is set to 40.

Refinement

To try to understand which hyperparameters would have been the best several trainings have been done with different settings with a subset of samples (200). Each training has been changed the batch size and the number of the epochs then has been evaluated the execution time and the performance.
Following an extract of the attempts:

2D Layers CNN & Pool

Epochs: 25	batch: 2	samples: 200	result: 14%	Wall time: 52min 44s
Epochs: 25	batch: 4	samples: 200	result: 14%	
Epochs: 25	batch: 16	samples: 200	result: 14%	Wall time: 1h 3min 11s
Epochs: 50	batch: 16	samples: 200	result: 20%	Wall time: 3h 1min 16s

4 2D CNN Convolutional Layer

Epochs: 5	batch: 8	samples: 200	result: 22%
Epochs: 13	batch: 8	samples: 200	result: 18%
Epochs: 14	batch: 8	samples: 200	result: 28%
Epochs: 16	batch: 8	samples: 200	result: 22%
Epochs: 30	batch: 8	samples: 200	result: 18%
Epochs: 14	batch: 4	samples: 200	result: 30%
Epochs: 14	batch: 4	samples: all	result: 25,83%
Epochs: 20	batch: 16	samples: all	result: 41,67%
Epochs: 25	batch: 16	samples: all	result: 40.83%
Epochs: 30	batch: 16	samples: all	result: 38,61%
Epochs: 50	batch: 16	samples: all	result: 47,50%

From this sample, training has been decided to train the model for the all set setting a batch size of 16 for 50 epochs.

Results

Model Evaluation and Validation

The final results of the models are the following:

Benchmark (linear):

Test Accuracy of angry: 32% (14/43)
Test Accuracy of surprise: 0% (0/45)
Test Accuracy of fearful: 0% (0/47)
Test Accuracy of neutral: 0% (0/51)
Test Accuracy of calm: 0% (0/54)
Test Accuracy of happy: 37% (15/40)
Test Accuracy of sad: 0% (0/33)
Test Accuracy of disgust: 0% (0/47)
Test Total Accuracy 8%

- Low total accuracy
- Most of the emotions are not predicted at all
- Two peaks for the “angry” and “happy” value

2D Layers CNN & Pool (Pytorch):

Test Accuracy of angry: N/A
Test Accuracy of surprise: 20% (0/45)
Test Accuracy of fearful: 61% (0/47)
Test Accuracy of neutral: 1% (0/51)
Test Accuracy of calm: 16% (0/54)
Test Accuracy of happy: 10% (15/40)
Test Accuracy of sad: 9% (0/33)
Test Accuracy of disgust: 2% (0/47)
Test Total Accuracy 9%

- Low total accuracy
- Good prediction on fearful
- More balanced predictions
- No angry sample in the test

4 2D CNN Convolutional Layer (Keras):

Test Accuracy of angry: 70% (24/34)
Test Accuracy of surprise: 17% (8/45)
Test Accuracy of fearful: 62% (34/54)
Test Accuracy of neutral: 33% (11/33)
Test Accuracy of calm: 72% (38/47)
Test Accuracy of happy: 29% (15/51)
Test Accuracy of sad: 34% (16/47)
Test Accuracy of disgust: 62% (25/40)
Test Total Accuracy 47.50%

- Higher accuracy
- Good prediction on fearful, calm and disgust
- Minor performance on surprise

Justification

Compare to the benchmark solution both algorithms work better.
The “2D Layers CNN & Pool” doesn’t have a much higher level of accuracy but the correction of the predictions is more spread through the classification with good accuracy about “fearful”.
The “*4 2D CNN Convolutional* Layer” has a good level of general prediction in with the highest peaks on “angry”, “fearful”,” calm” and “disgust” emotion. The model doesn’t reach the same accuracy of the original test, this probably because the data set is smaller.

The combination of the two results shows that with this procedure (MEL spectrum) there is a good potential to recognize the “fearful” sentiment.

Conclusions

In this project, it has been tested two models for the speech emotion recognition analysis.
No of the two models give a significant result in order to be utilized in a professional field but they mark a good starting point.

The first case has been tested the *2D Layers CNN & Pool* model that gave a 9% accuracy, 1% more than the benchmark model but with more spread correct predictions.
In the second case the *4 2D CNN Convolutional* model, reached 47.50% accuracy with over 70% of the right prediction for calm and angry, while is underperforming the previous model just regard the “surprise” prediction.

So, the project shows that the second model is bringing more accuracy, it underperforms the original test probably because the dataset used is smaller, so, probably there are rooms of improvement feeding it with more data.

For the next steps is possible to think to try to improve the first one training it with more data, as well for the second one that can be also improve trying to add more convolution layers.
Finally, if the results are significant and different in performing is possible to try to combine and weigh the two model predictions in order to increase the total accuracy.