Emotion Recognition

Speech Processing and Speech Technologies

Edgar Andrés Santamaría

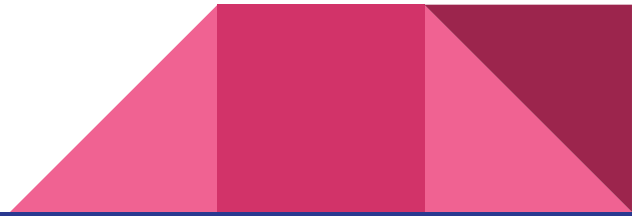
eandres011@ikasle.ehu.eus

The Introduction

The task and some possible applications

The task consists in the automatic detection of the emisor emotions in certain Speech.

This technology could be used in customer satisfaction application purposes for example in call centers.



The Resources

Data

ravdess-emotional-song/speech-audio:

- 44 trials per actor x 24 actors = 1012 (.wav) files. (Song)
- 60 trials per actor x 24 actors = 1440 (.wav) files. (Speech)
- Emotions includes calm, happy, sad, angry, and fearful expressions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression.

Data

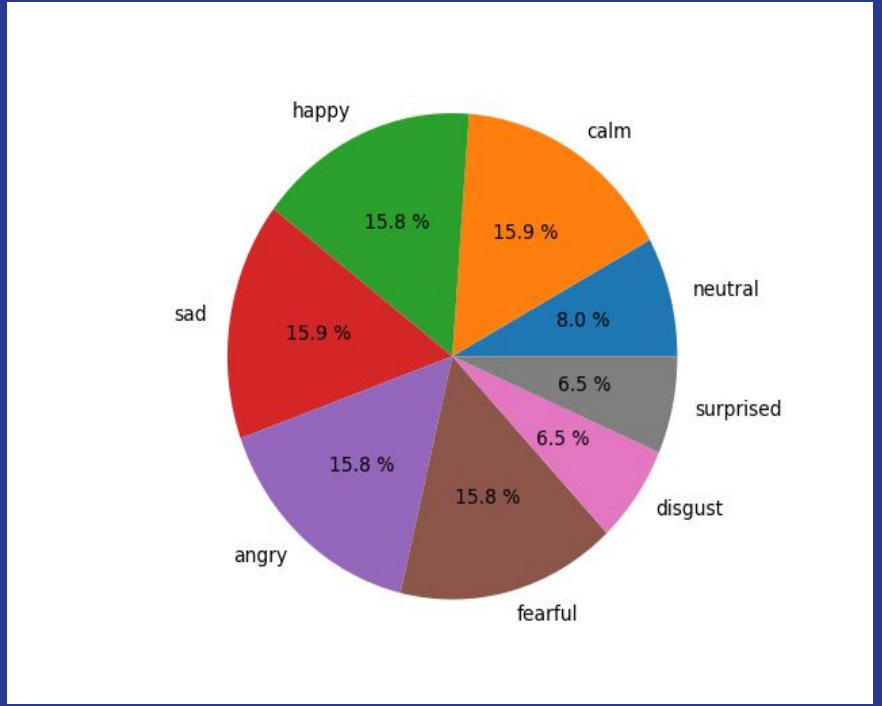
ravdess-emotional-song/speech-audio:

- 24 professional actors (12 female, 12 male).
- vocalizing two lexically-matched statements in a neutral North American accent.
- Statements: "Kids are talking by the door", and "Dogs are sitting by the door".
- total data 2452 (.wav) files.

Data

Train Distribution

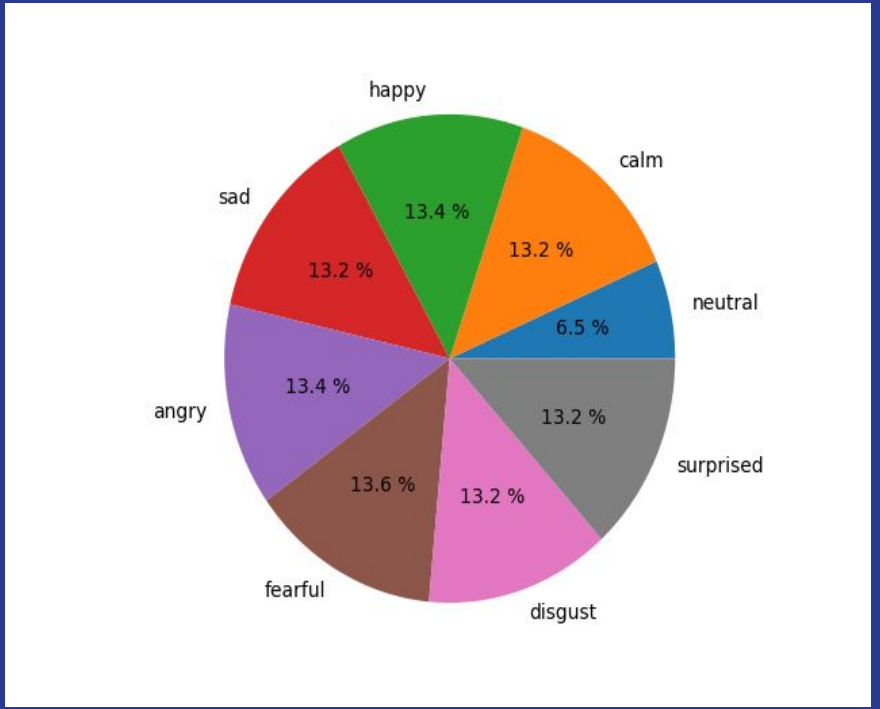
2452 * 0.8 = 1962



Data

Test Distribution

2452 * 0.2 = 490

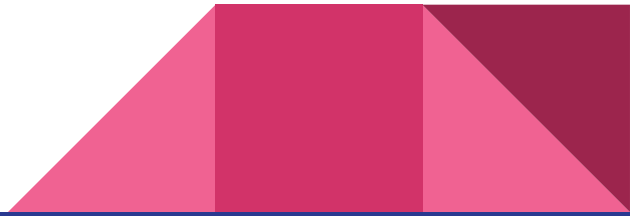


The Challenge

Objectives

We define the requirements of the system:

1. Avoids the silence.
2. Enhances the sound properties.
3. Takes care of vocal tract.
4. Takes care of the message.
5. generalizes the detection.
6. Handles multi labeling task.
7. Reports properly the evaluation.
8. Reports properly the training process.



Data preprocess

Here we aim to improve the data quality :

1. Trim : Consists on quitting automatically those parts of the sound wave with less power than a threshold to avoid silence.

silence threshold 30 DB

Tools: (open source Software)

1. Librosa

Emotion Recognition

The task as it's proposed requires a lot of generalization capacity, due we feed only two messages and the matter is to analyze the vocal tract properly, and the intensity of the message.

For this reason we propose the [Mel Spectrogram (2D) ; Chroma (2D) ; Tempogram (2D)] information as 3D input for the big Deep model, and Mel [Spectrogram (1D) ; Chroma (1D) ; Tempogram (1D)] mean coefficients as 2D input for the baseline.

Data representation

Here we aim to represent the data in a way that deep learning approach could learn over: (input 3-D data)

1. MFCCS
2. Chroma
3. Tempogram
4. Mel Spectrogram

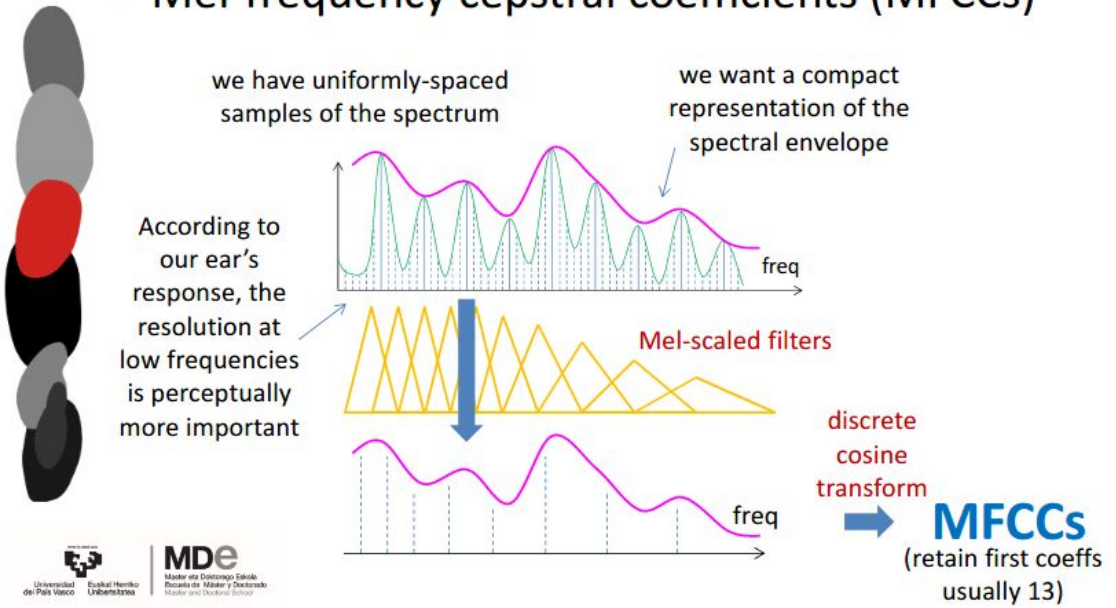
The example :

Actor: Male
Tag: Neutral
Sentence: Kids are talking by the door
prediction: neutral

Tools: (open source Software)

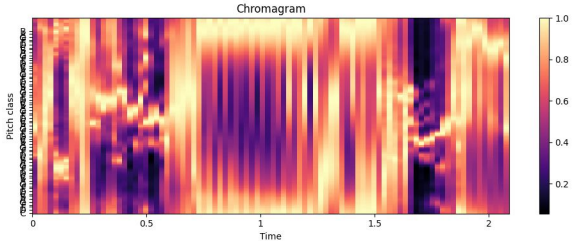
1. Librosa
2. Matplotlib

Mel-frequency cepstral coefficients (Mfccs)



In this case we also want to track the excitement, so 40 coefficients were used.

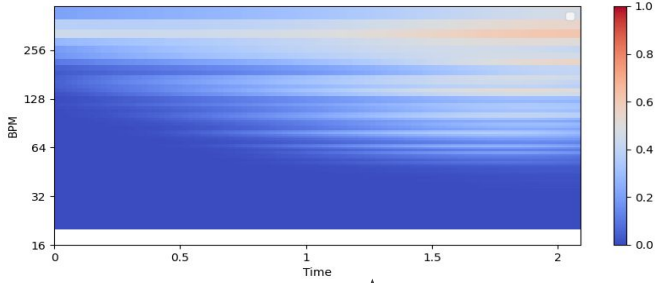
Chroma



use the simpler procedure of mapping each STFT bin directly to chroma, after selecting only spectral peaks. (Coeffs)

The chroma representation tells us the intensity of each of the 90 distinct chroma at each time frame. could be turned back into signal simply by using the sinusoids. (Matrix)

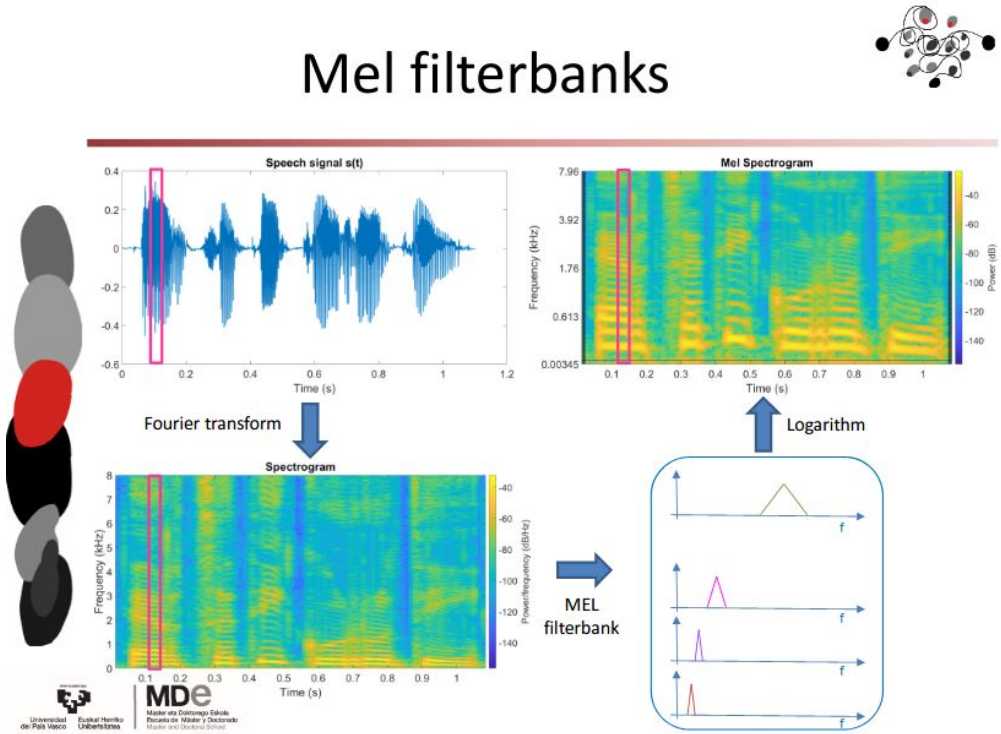
Tempogram



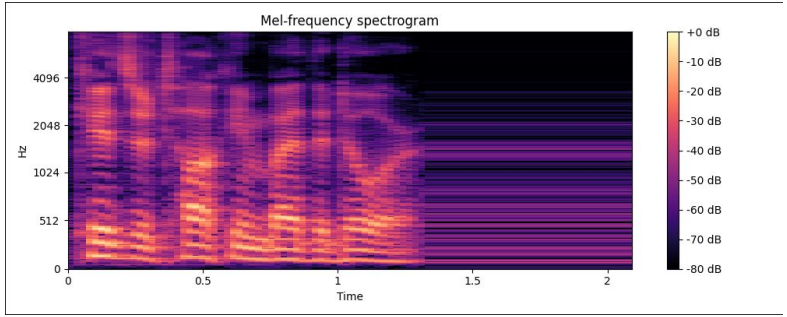
First the Fourier transform is computed, the Spectrogram

Then we calculate the beats per minute (BPM) for each frame, all the process is normalized to the tempo parameter, in this case 128.

Mel Spectrogram



512 Mel Filters
fixed into 90 length
padding ‘edge’

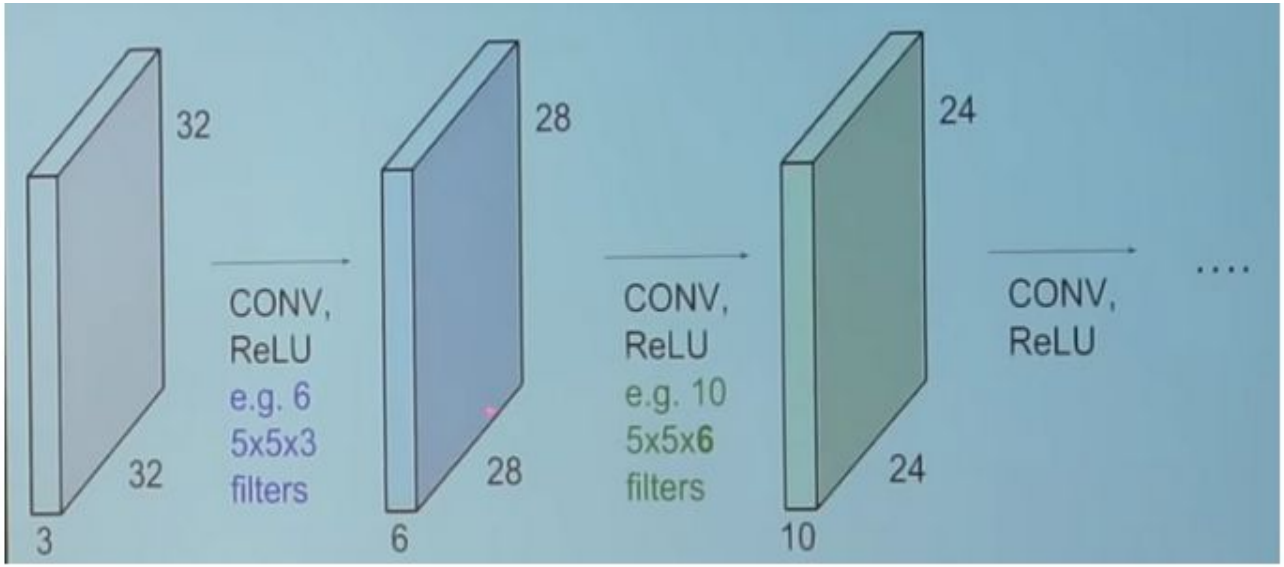


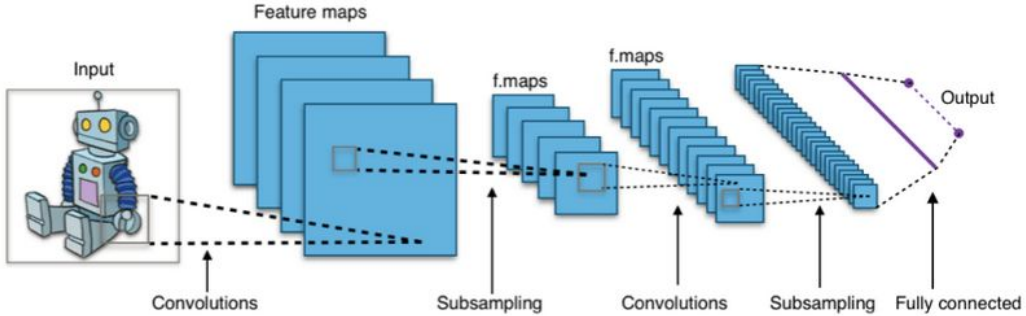
Deep model approach

This technology allows us reporting properly over training steps and feedback the evaluation in the way we want.

Has also some powerful features in the models called 2D convolutional neural networks (CNNs) that consists on filtering layers, those aim to gather the correct input at every stage (performs dimensional reduction) until we have the correct data to provide accurate outputs.

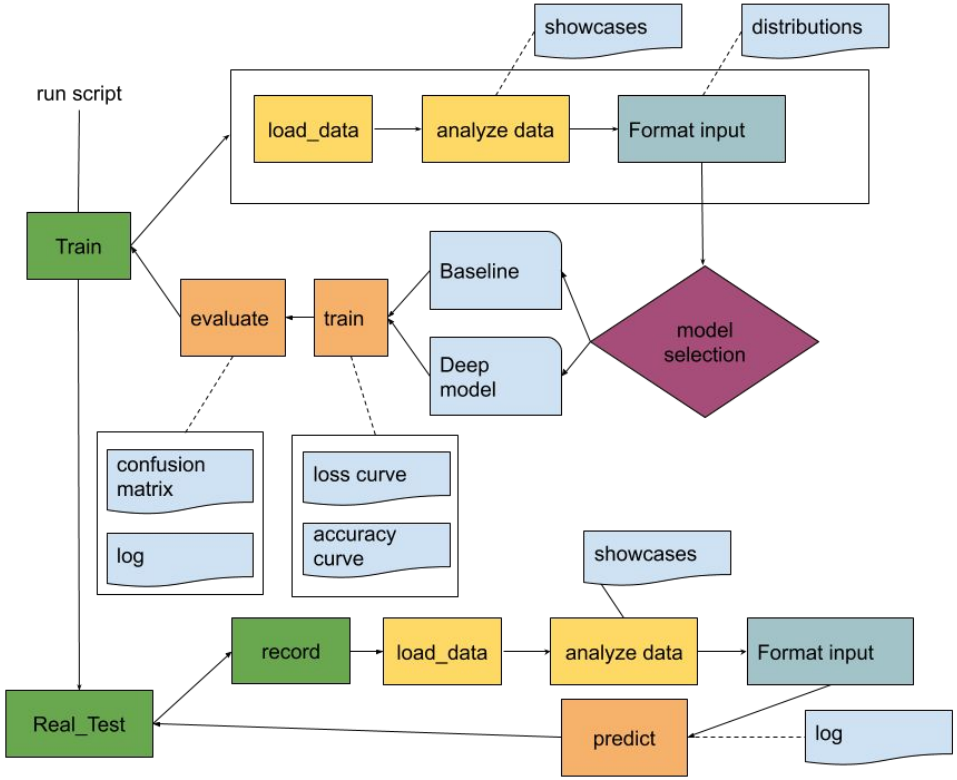
CNNs


32
32
3
CONV,
ReLU
e.g. 6
5x5x3
filters
28
28
6
CONV,
ReLU
e.g. 10
5x5x6
filters
24
24
10
CONV,
ReLU
....


Feature maps
Input
f.maps
f.maps
Output
Convolutions
Subsampling
Convolutions
Subsampling
Fully connected

The Experiment

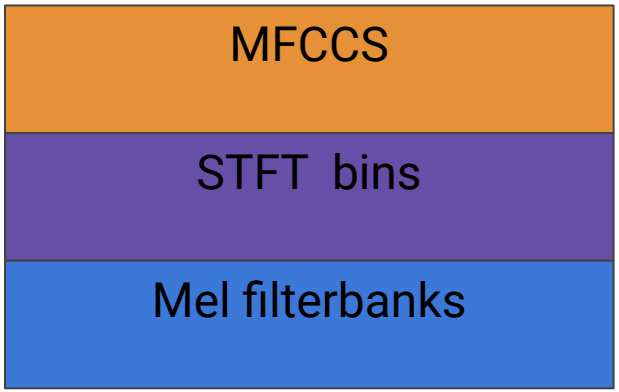
Design


showcases
distributions
run script
load_data
analyze data
Format input
Train
Baseline
evaluate
train
model selection
Deep model
confusion matrix
loss curve
log
accuracy curve
showcases
record
load_data
analyze data
Format input
Real_Test
predict
log

Feature extraction - Baseline

Baseline uses fully connected layer that requires matrix input (2D).

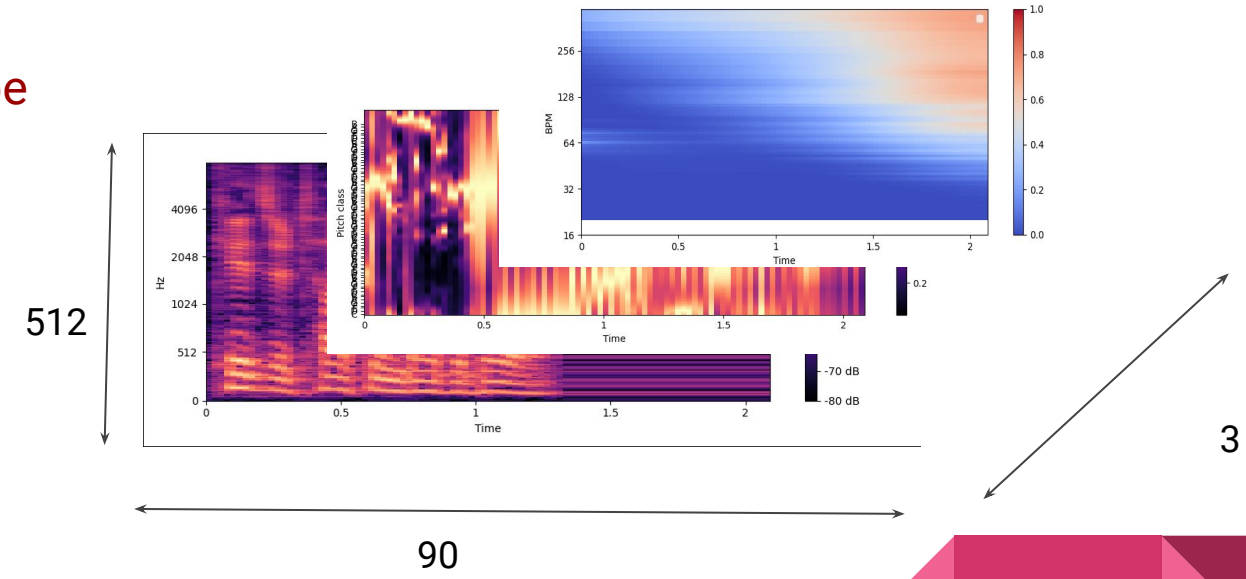
Input is based on the means of MFCCs, chroma STFT and Mel spectrogram.



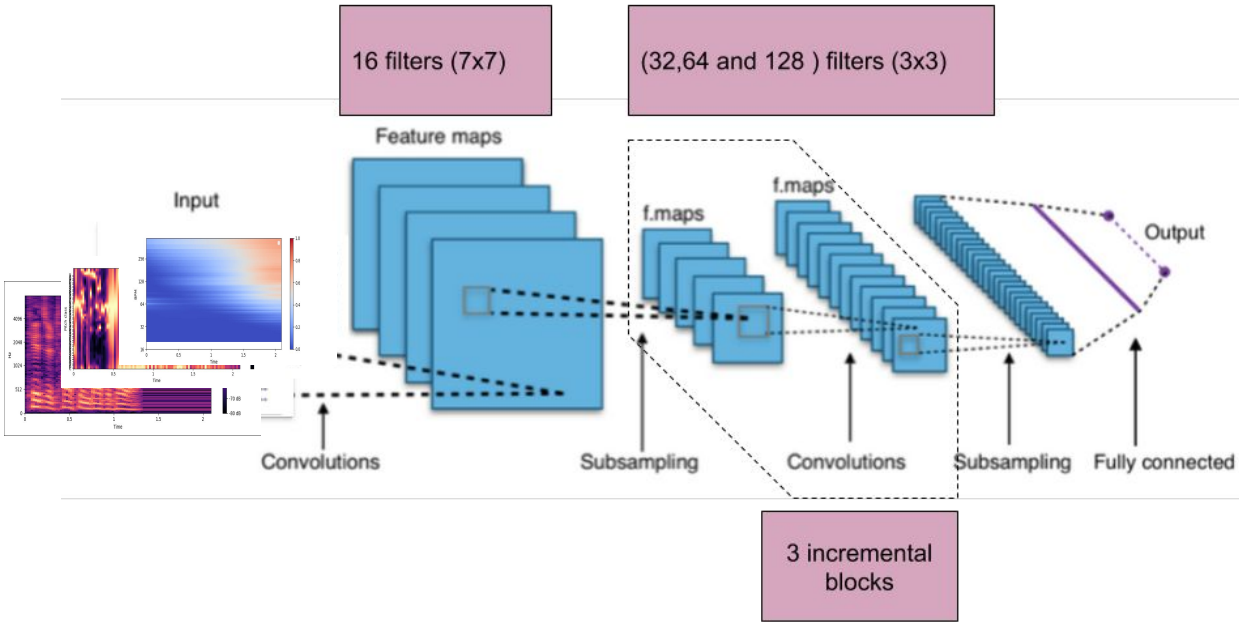
Feature extraction - Deep Model

Deep model uses 2D CNN that requires cube input (3D).

this input is based on Mel Spectrogram, Chroma and Tempogram.



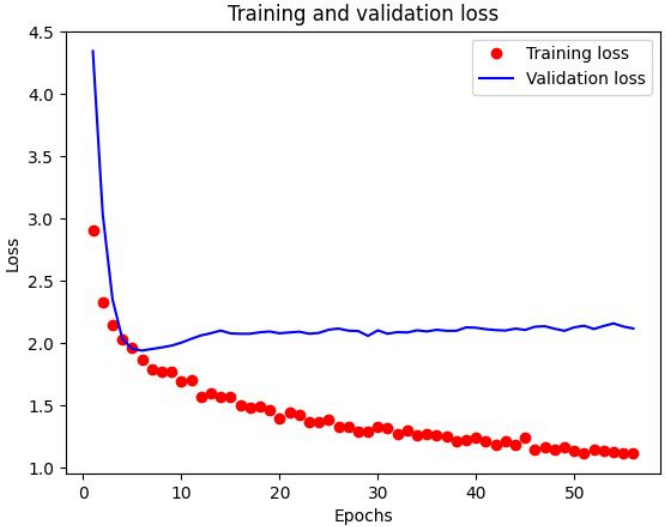
Architecture - Deep model

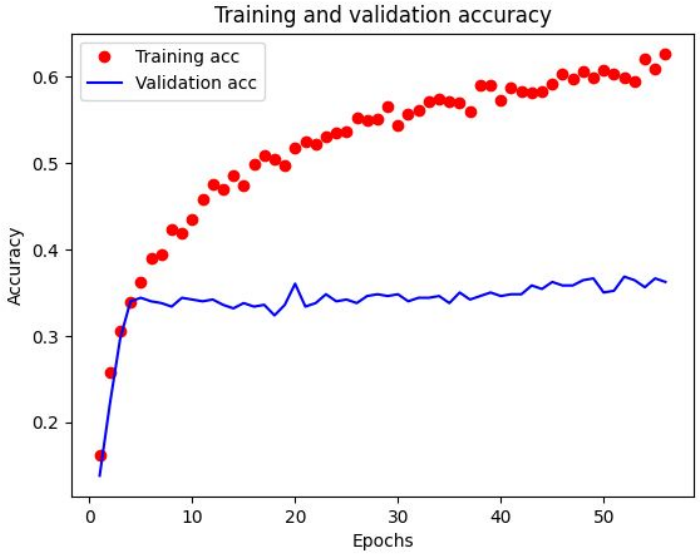


The architecture proposed in this case uses 2D CNNs to filter the input using 3 blocks.

The Conclusions

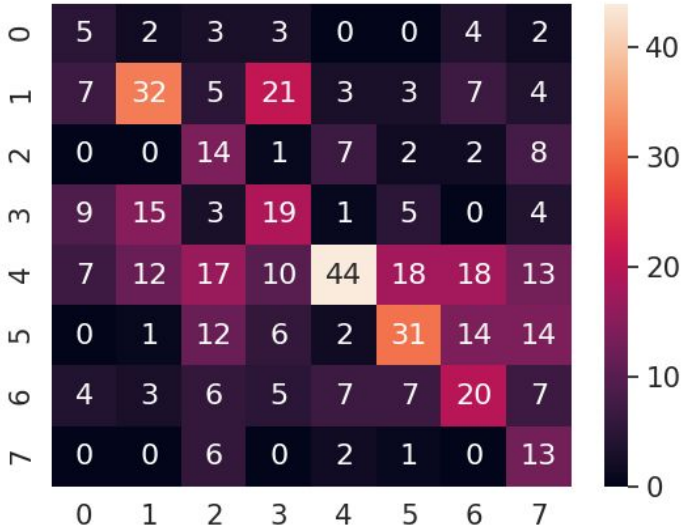
Results - Baseline Model



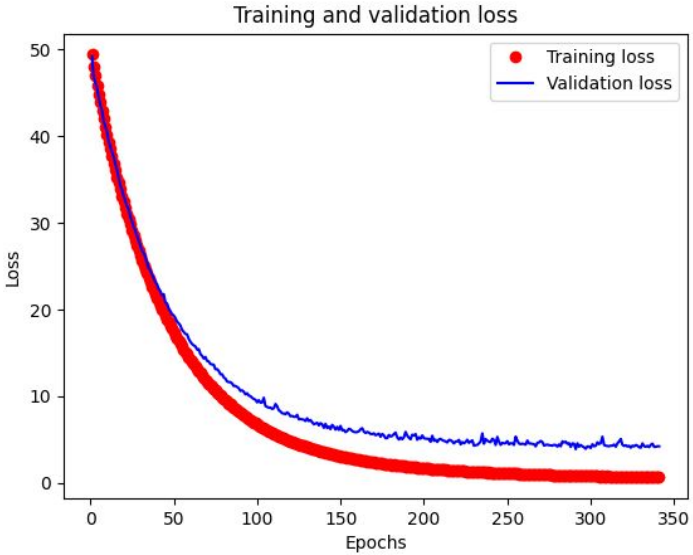


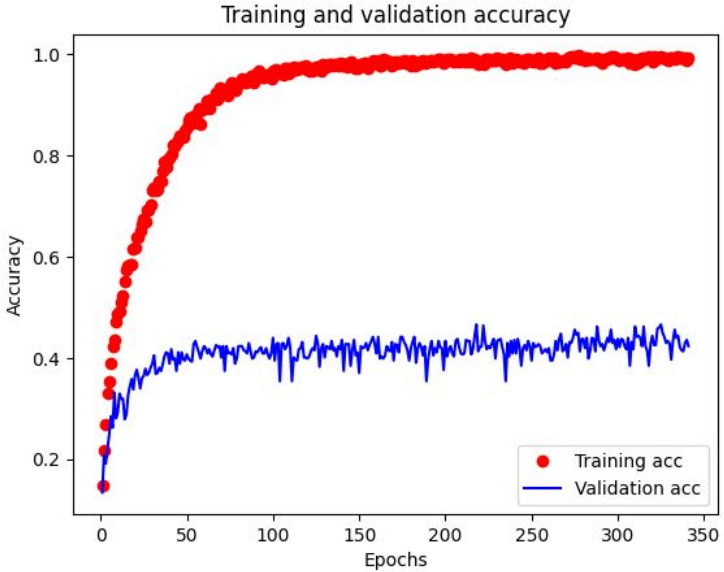
Results - Baseline Model

```
[INFO] evaluating network...
              precision    recall  f1-score   support

     neutral       0.26      0.16      0.20        32
        calm       0.39      0.49      0.44        65
       happy       0.41      0.21      0.28        66
         sad       0.34      0.29      0.31        65
       angry       0.32      0.67      0.43        66
     fearful       0.39      0.46      0.42        67
     disgust       0.34      0.31      0.32        65
   surprised       0.59      0.20      0.30        65

    accuracy                           0.36       491
   macro avg       0.38      0.35      0.34       491
weighted avg       0.39      0.36      0.35       491


Process finished with exit code 0
```



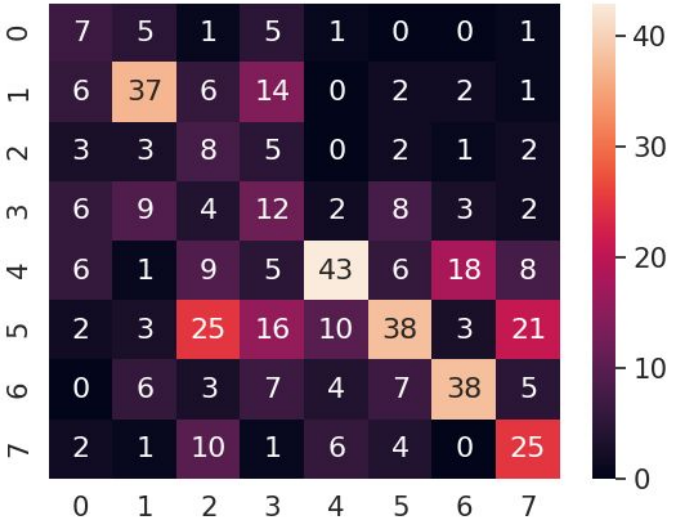
Results - Deep Model





Results - Deep Model

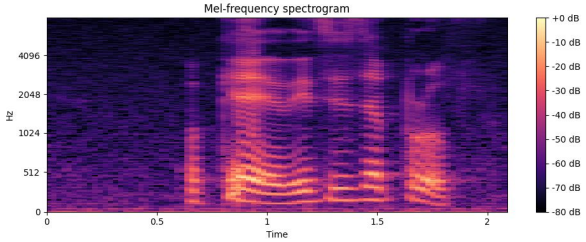
```
[INFO] evaluating network...
              precision    recall  f1-score   support

     neutral       0.35      0.22      0.27        32
        calm       0.54      0.57      0.56        65
       happy       0.33      0.12      0.18        66
         sad       0.26      0.18      0.22        65
       angry       0.45      0.65      0.53        66
     fearful       0.32      0.57      0.41        67
     disgust       0.54      0.58      0.56        65
   surprised       0.51      0.38      0.44        65

    accuracy                           0.42       491
   macro avg       0.41      0.41      0.40       491
weighted avg       0.42      0.42      0.40       491


Process finished with exit code 0
```

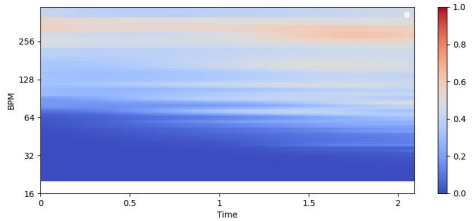


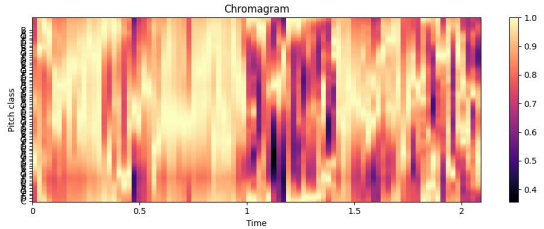
Practical testing 1

First of all notice that the subject doesn't necessarily have to discover it's emotions.

Happy



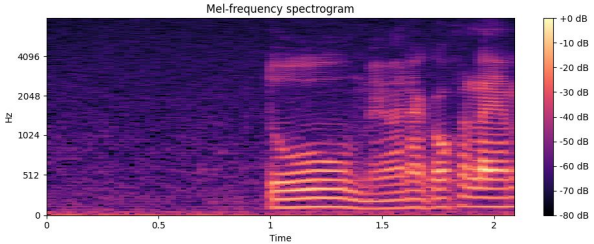


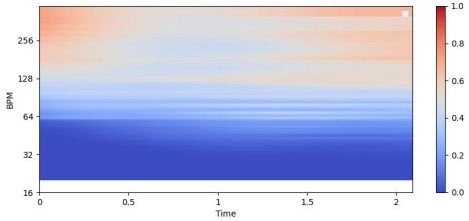


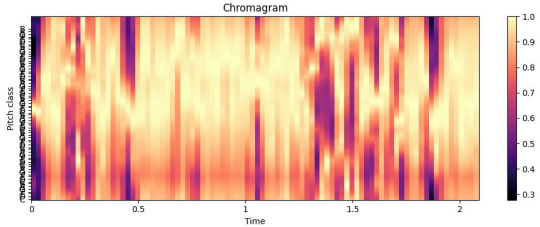
Practical testing 2

Finally notice that the subject doesn't have to speak english.

calm







Conclusions

We conclude that :

1. The approach fits quality requirements.
2. The approach could be improved (new Architecture even features).
3. The approach constitutes the first methodological approach to the task.
4. Convolutional 2D layers hold well the problem.
5. In this task the whole matter consists on the representation of the signal.
6. We ensure that results of the model are achieved practically.

https://github.com/EdgarAndresSantamaria/Speech_Emotion_Recognition

Bibliography

- Venkataramanan, K., & Rajamohan, H. R. (2019). Emotion Recognition from Speech. arXiv preprint arXiv:1912.10458.
- https://www.pyimagesearch.com/2018/12/31/keras-conv2d-and-convolutional-layers/
- Thoma, M. (2017). Analysis and optimization of convolutional neural network architectures. arXiv preprint arXiv:1707.09725.

Bibliography

- Lim, W., Jang, D., & Lee, T. (2016, December). Speech emotion recognition using convolutional and recurrent neural networks. In 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA) (pp. 1-4). IEEE.
- Kuchibhotla, S. , Vankayalapati, H. , Vaddi, R. , Anne, K.R. , 2014. A comparative analysis of classifiers in emotion recognition through acoustic features. Int. J. Speech Technol. 17 (4), 401–408 .

Bibliography

- Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., & Shaalan, K. (2019). Speech recognition using deep neural networks: A systematic review. IEEE Access, 7, 19143-19165.
- Akçay, M. B., & Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication, 116.

Bibliography

- Zen, H., Tokuda, K., & Black, A. W. (2009). Statistical parametric speech synthesis. speech communication, 51(11), 1039-1064.
- Grosche, P., Müller, M., & Kurth, F. (2010, March). Cyclic tempogram a mid-level tempo representation for music signals. In 2010 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 5522-5525). IEEE.

Bibliography

- Ellis, Daniel P.W. “Chroma feature analysis and synthesis” 2007/04/21 http://labrosa.ee.columbia.edu/matlab/chroma-ansyn/.

Questions :)