Emotion Detection Using Speech

Zeduo Zhang
Kun Wang

Overview

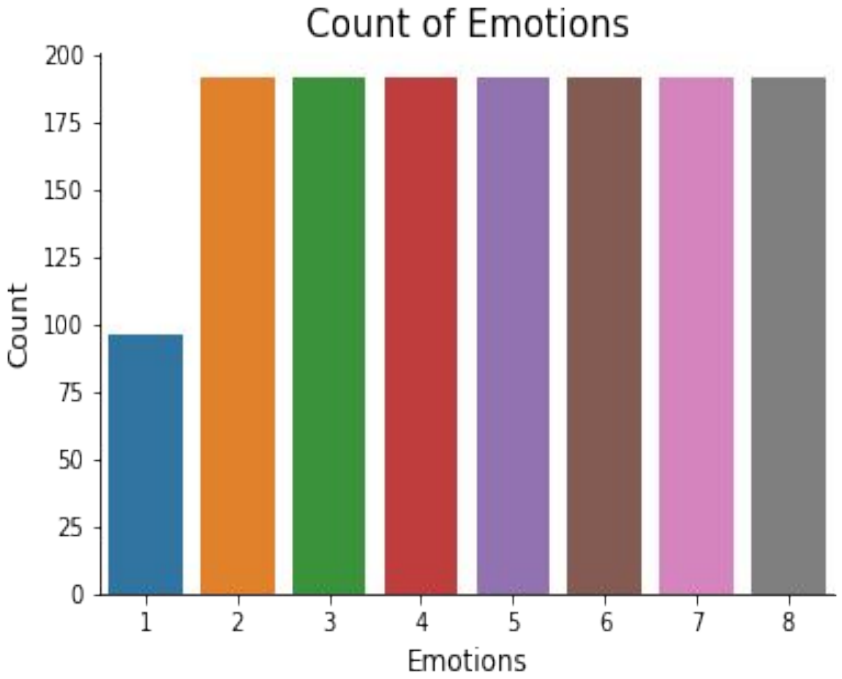
1. Motivation & Objective
2. Datasets
3. Features
4. Data Augmentation
5. Learning rate
6. Results
7. Future works

Motivation & Objective

Study on how to recognize the different emotions from speech.

Achieve a high accuracy on emotion detection.

Datasets



1- Neutral

2- Calm

3- Happy

4- Sad

5- Angry

6- Fearful

7- Disgust

8- Surprised

Datasets

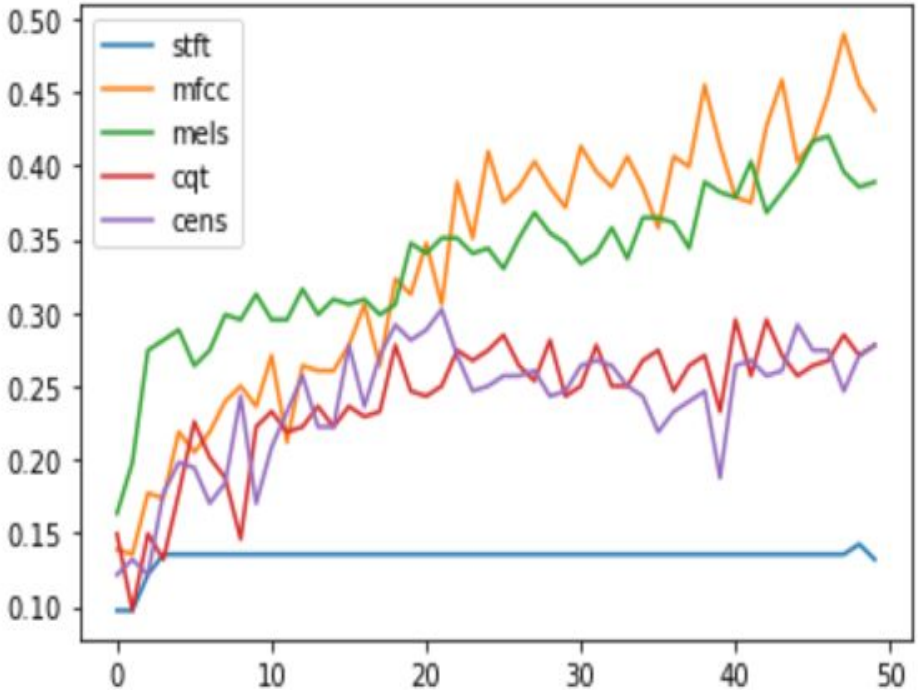
	Url	Modality	Vocal channel	Emotion	Emotional intensity	Statement	Repetition	Actor
0	03-01-01-01-01-01-01.wav	3	1	1	1	1	1	1
1	03-01-01-01-01-02-01.wav	3	1	1	1	1	2	1
2	03-01-01-01-02-01-01.wav	3	1	1	1	2	1	1
3	03-01-01-01-02-02-01.wav	3	1	1	1	2	2	1
4	03-01-02-01-01-01-01.wav	3	1	2	1	1	1	1
...	...	...	...	...	...	...	...	...
1435	03-01-08-01-02-02-24.wav	3	1	8	1	2	2	24
1436	03-01-08-02-01-01-24.wav	3	1	8	2	1	1	24
1437	03-01-08-02-01-02-24.wav	3	1	8	2	1	2	24
1438	03-01-08-02-02-01-24.wav	3	1	8	2	2	1	24
1439	03-01-08-02-02-02-24.wav	3	1	8	2	2	2	24

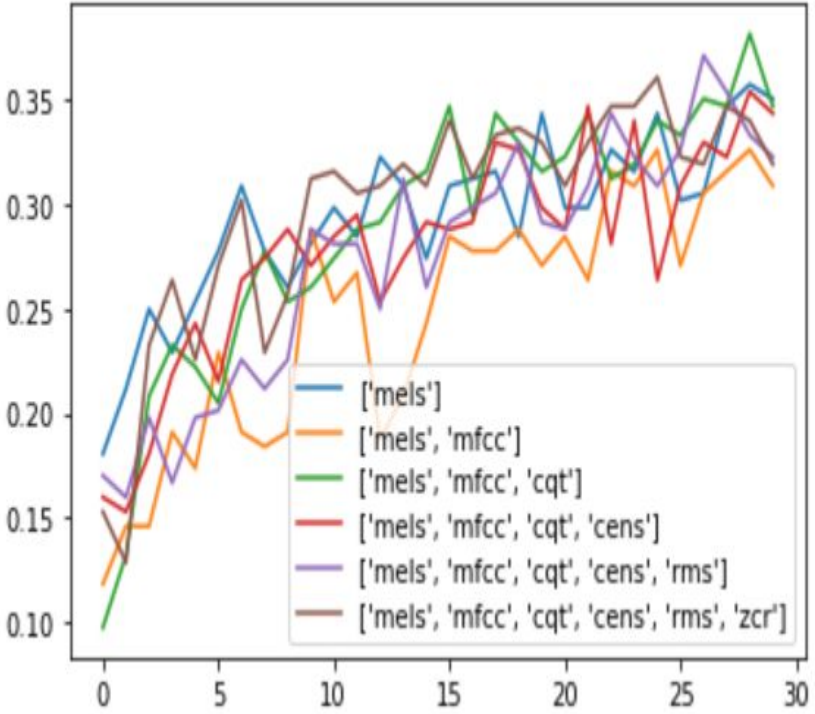
1440 rows × 8 columns

Datasets

- Modality (01 = full-AV, 02 = video-only, 03 = audio-only).
- Vocal channel (01 = speech, 02 = song).
- Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).
- Emotional intensity (01 = normal, 02 = strong).
- Statement (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door").
- Actor (01 to 24. Odd numbered actors are male, even numbered actors are female).

Features


stft
mfcc
mels
cqt
cens
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0
10
20
30
40
50


['mels']
['mels', 'mfcc']
['mels', 'mfcc', 'cqt']
['mels', 'mfcc', 'cqt', 'cens']
['mels', 'mfcc', 'cqt', 'cens', 'rms']
['mels', 'mfcc', 'cqt', 'cens', 'rms', 'zcr']
0.35
0.30
0.25
0.20
0.15
0.10
0
5
10
15
20
25
30

MFCCs (Mel-frequency cepstrum coefficients)

MFCCs of a signal describe the overall shape of a spectral envelope

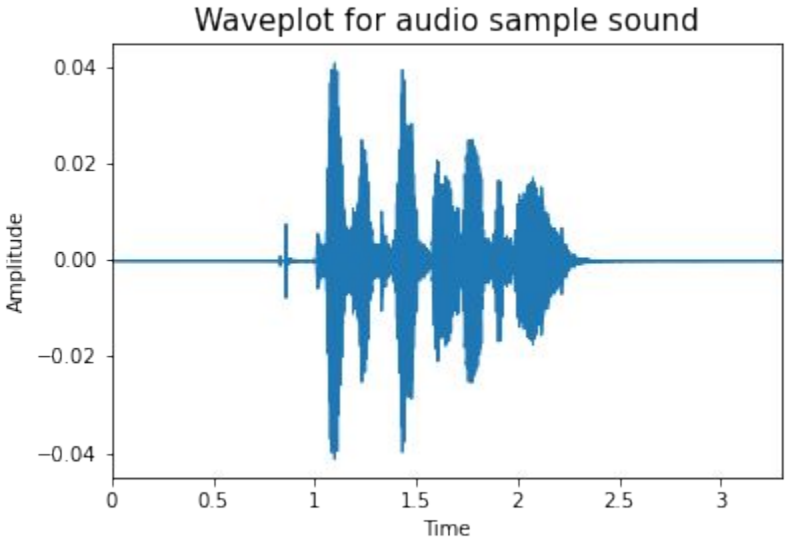
MFCCs are commonly used as features in speech recognition systems, such as the systems which can automatically recognize numbers spoken into a telephone

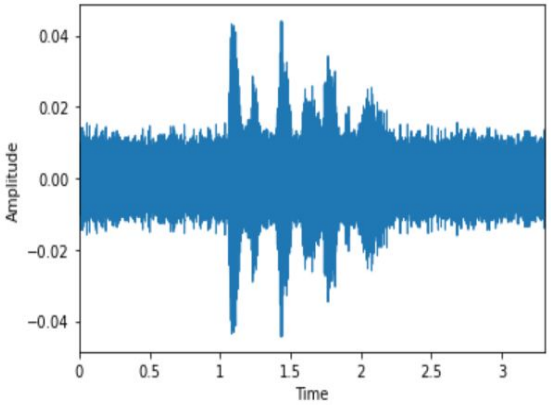
Data Augmentation

- Noise
- Time shifting
- Time stretching
- Pitch shifting

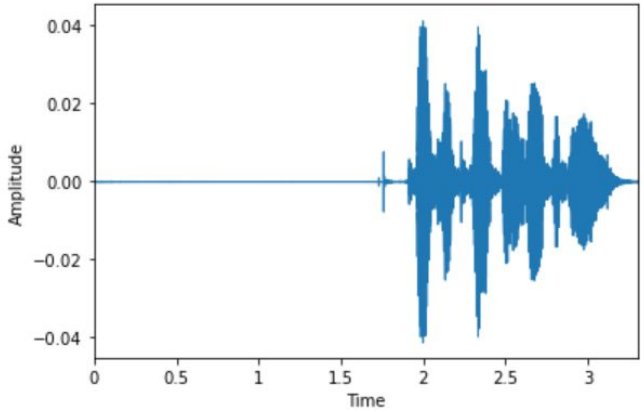
Data augmentation is a method for generating synthetic data

Neutral

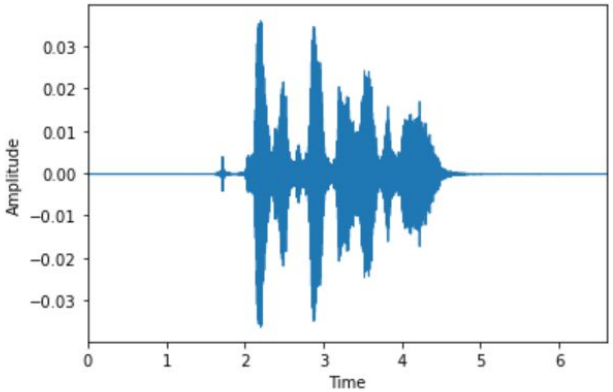

Waveplot for audio sample sound
Amplitude
0.04
0.02
0.00
−0.02
−0.04
0
0.5
1
1.5
2
2.5
3
Time



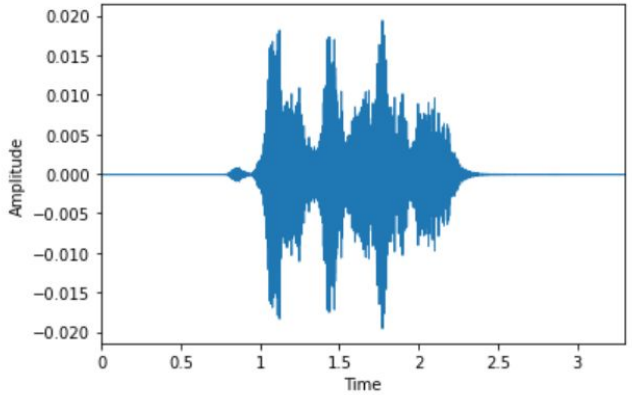
Noise



Time Shifting



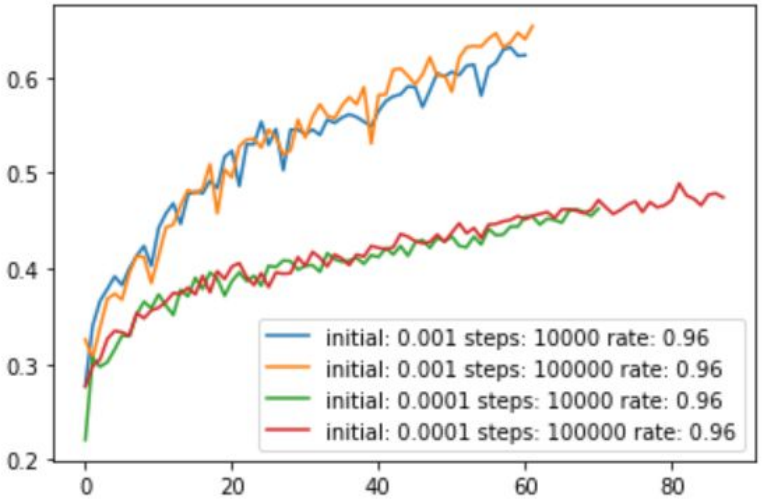
Time Stretching

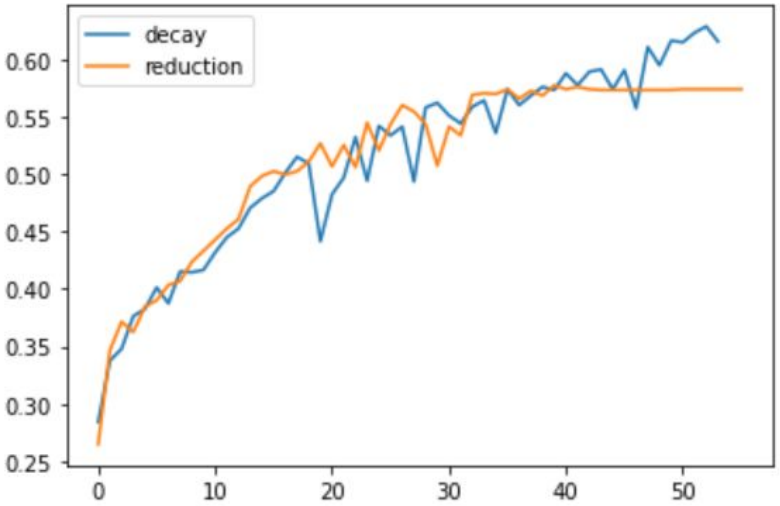


Pitch Shifting

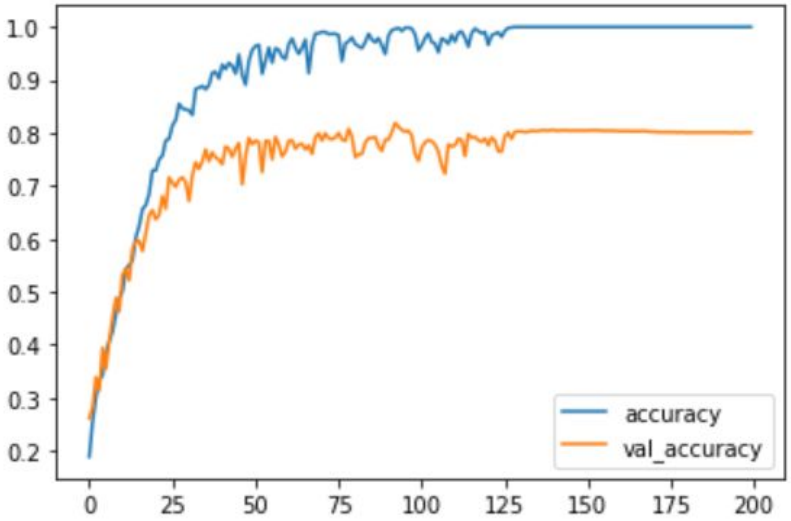
Learning rate

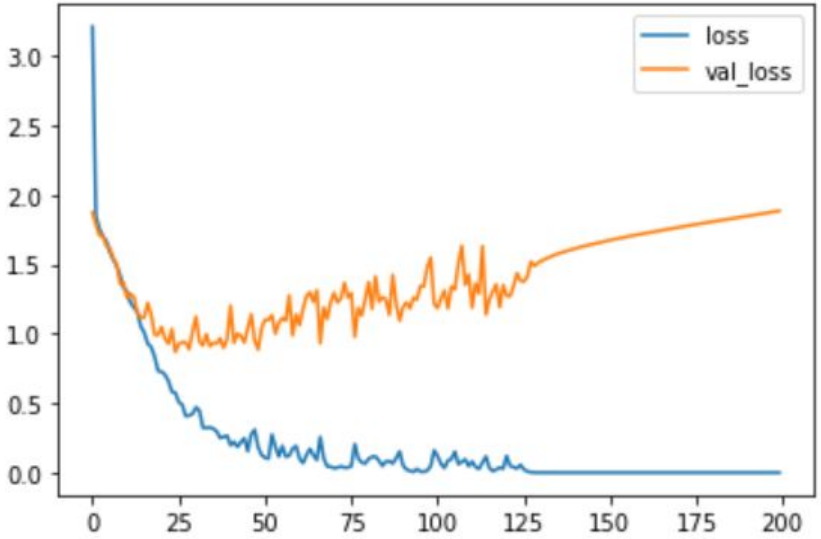
- Decay
- Reduction





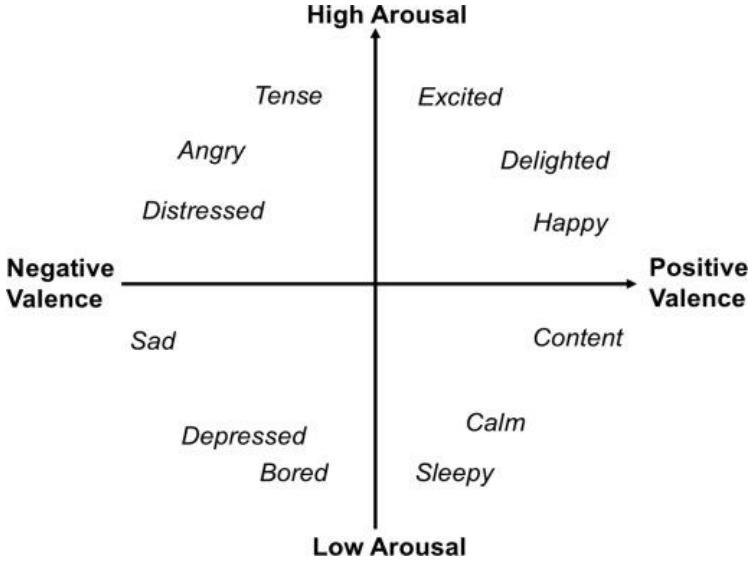
Results


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0
25
50
75
100
125
150
175
200
accuracy
val_accuracy


3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
25
50
75
100
125
150
175
200
loss
val_loss

Future works

- Run all the models and comparison on a larger number of epochs
- Add more emotional speech dataset to increase the diversity of our dataset
- Try to find a way to evaluate the arousal and valence of an audio, and recognize the emotion based on these two characteristics.



Reference

https://towardsdatascience.com/regularization-an-important-concept-in-machine-learning-5891628907ea

https://towardsdatascience.com/a-practical-introduction-to-early-stopping-in-machine-learning-550ac88bc8fd

https://www.sciencedirect.com/science/article/pii/S0968090X19313099

https://www.youtube.com/watch?v=4_SH2nfbQZ8

https://en.wikipedia.org/wiki/Mel-frequency_cepstrum