Speech Emotion Recognition

Bohan Wang

Objective

The objective is to make a deep learning model which can classify speech into various categories of emotions.

The application of the speech emotion recognition system :

- psychiatric diagnosis
- intelligent toys
- Lie detection

1.Common Technique

Pipeline of speech emotion recognition

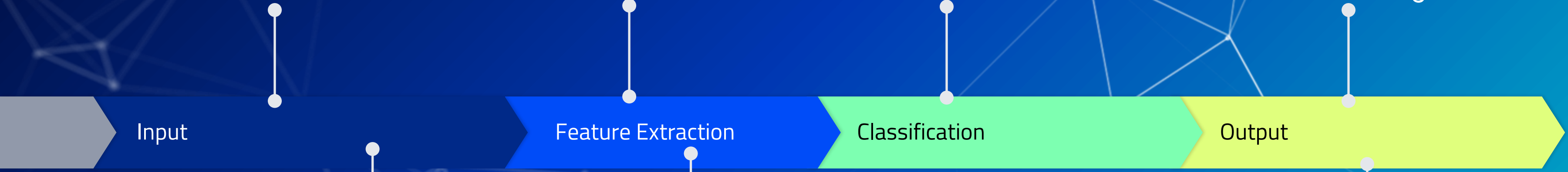
Pipeline

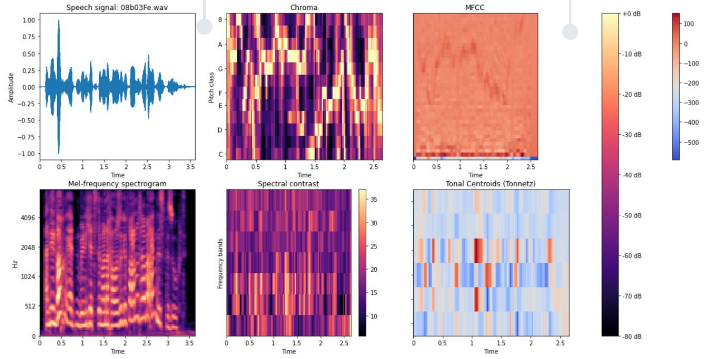
Audio files in .wav, .mp3 format

- □ MFCC (mfcc)
- □ Chroma (chroma)
- □ Mel Sepectrogram Frequency (mel)
- □ Contrast (contrast)
- □ Tonnetz (tonnetz)

CNN

The different emotion to which the audio files belong.





Speech Corpus: EmoDB

535 Utterances

The EMODB database is the freely available German emotional database. The database is created by the Institute of Communication Science, Technical University, Berlin, Germany. Ten professional speakers (five males and five females) participated in data recording. The database contains a total of 535 utterances.

7 emotions

The EMODB database comprises of seven emotions: 1) anger; 2) boredom; 3) anxiety; 4) happiness; 5) sadness; 6) disgust; and 7) neutral.

16 kHz sampling rate

The data was recorded at a 48-kHz sampling rate and then down-sampled to 16-kHz.

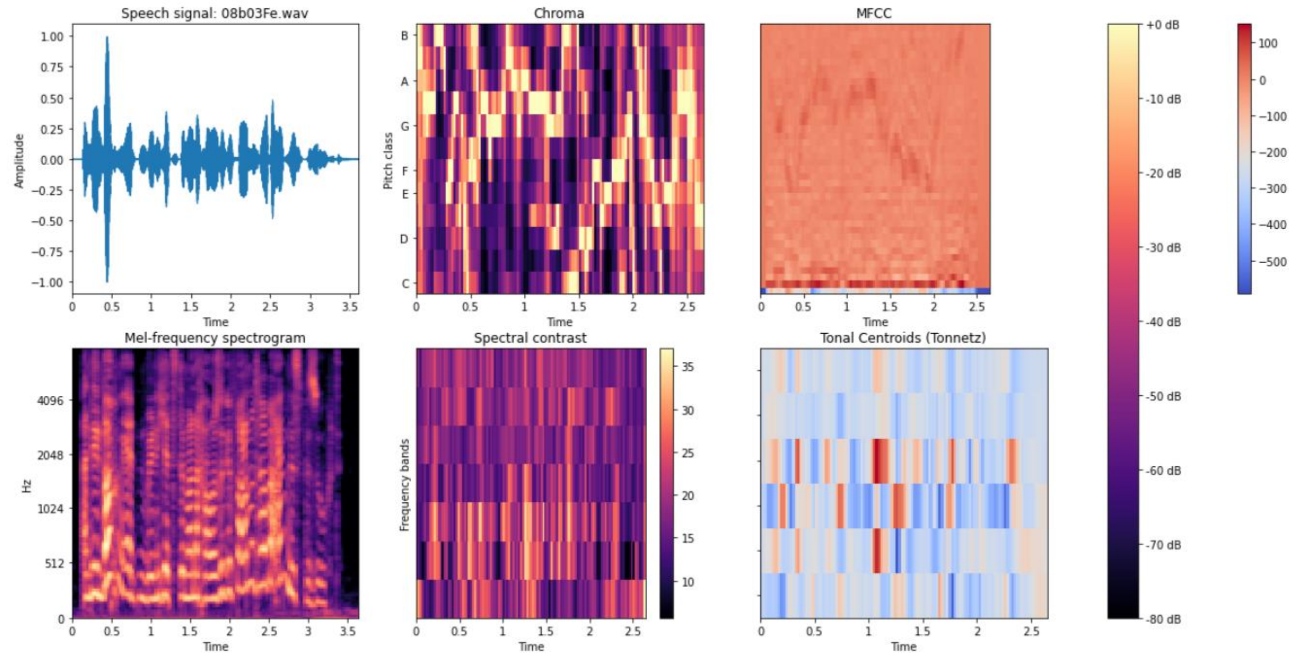
Data Augmentation

		Random	librosa function
1	Time-stretch an audio series by a random rate.	Uniformly distributed over [0.3, 2.0)	librosa.effects.time_stretch
2	Shift the pitch of a waveform by random steps	Uniformly distributed over [-2, 2.)	librosa.effects.pitch_shift

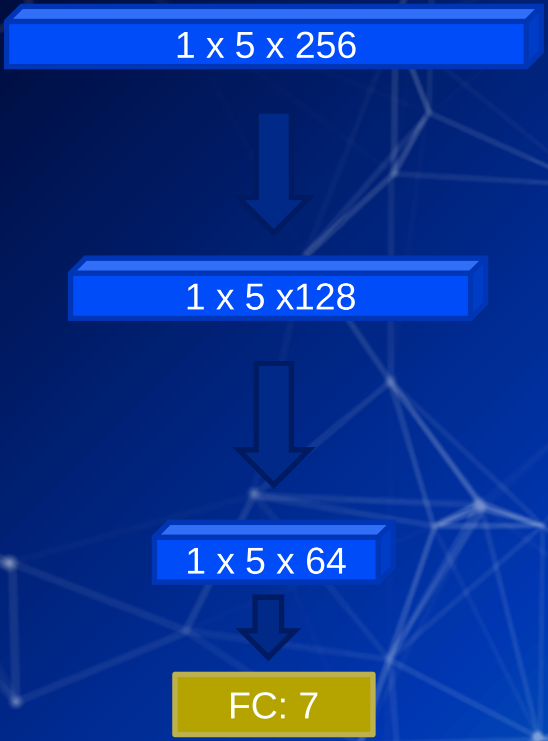
Feature Extraction

Features: Based on speech signal processing library (librosa)

1. MFCC (mfcc)
2. Chroma (chroma)
3. Mel Sepectrogram Frequency (mel)
4. Contrast (contrast)
5. Tonnetz (tonnetz)



CNN

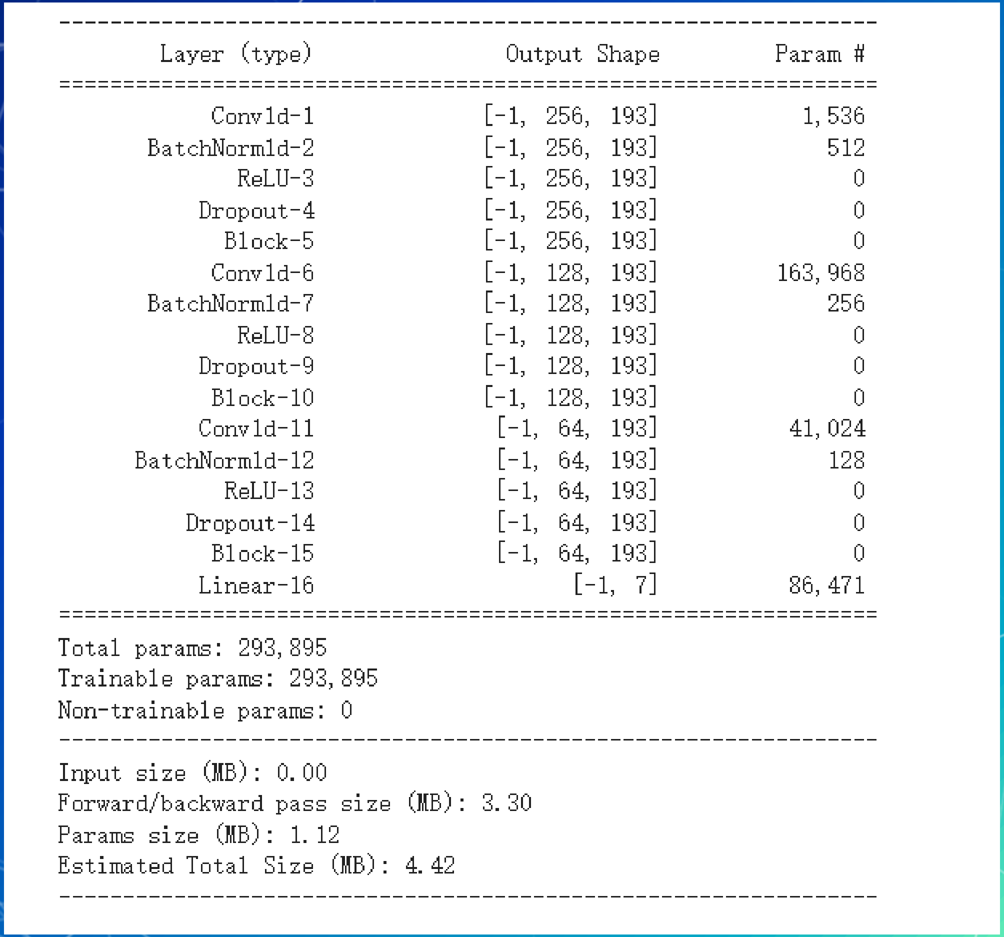


1 x 5 x 256

1 x 5 x128

1 x 5 x 64

FC: 7



```
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
            Conv1d-1              [-1, 256, 193]           1,536
       BatchNorm1d-2              [-1, 256, 193]             512
              ReLU-3              [-1, 256, 193]               0
           Dropout-4              [-1, 256, 193]               0
             Block-5              [-1, 256, 193]               0
            Conv1d-6              [-1, 128, 193]         163,968
       BatchNorm1d-7              [-1, 128, 193]             256
              ReLU-8              [-1, 128, 193]               0
           Dropout-9              [-1, 128, 193]               0
            Block-10              [-1, 128, 193]               0
           Conv1d-11               [-1, 64, 193]          41,024
      BatchNorm1d-12               [-1, 64, 193]             128
             ReLU-13               [-1, 64, 193]               0
          Dropout-14               [-1, 64, 193]               0
            Block-15               [-1, 64, 193]               0
           Linear-16                     [-1, 7]          86,471
================================================================
Total params: 293,895
Trainable params: 293,895
Non-trainable params: 0
----------------------------------------------------------------
Input size (MB): 0.00
Forward/backward pass size (MB): 3.30
Params size (MB): 1.12
Estimated Total Size (MB): 4.42
----------------------------------------------------------------
```

5 Fold cross validation

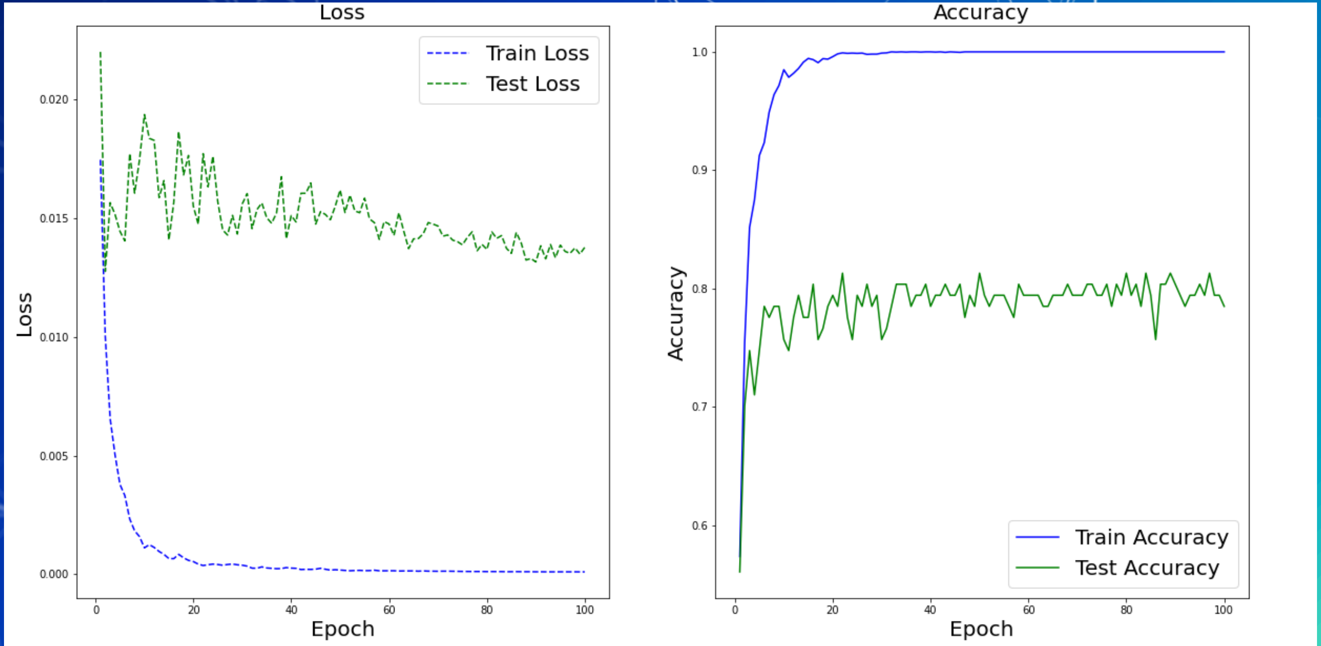
Best Hyperparameters		Value
1	Learning rate	0.001
2	Weight decay	0.005
3	gamma for exponential learning rate schedule	0.95

Training and Results

Training setup:

1. Epoch: 100
2. Batch size: 64
3. Optimizer: Adam
4. Learning rate: 0.001
5. Weight decay: 0.005
6. Gamma for exponential learning rate schedule: 0.95

Train accuracy: 100%
Best test accuracy: 81.3%

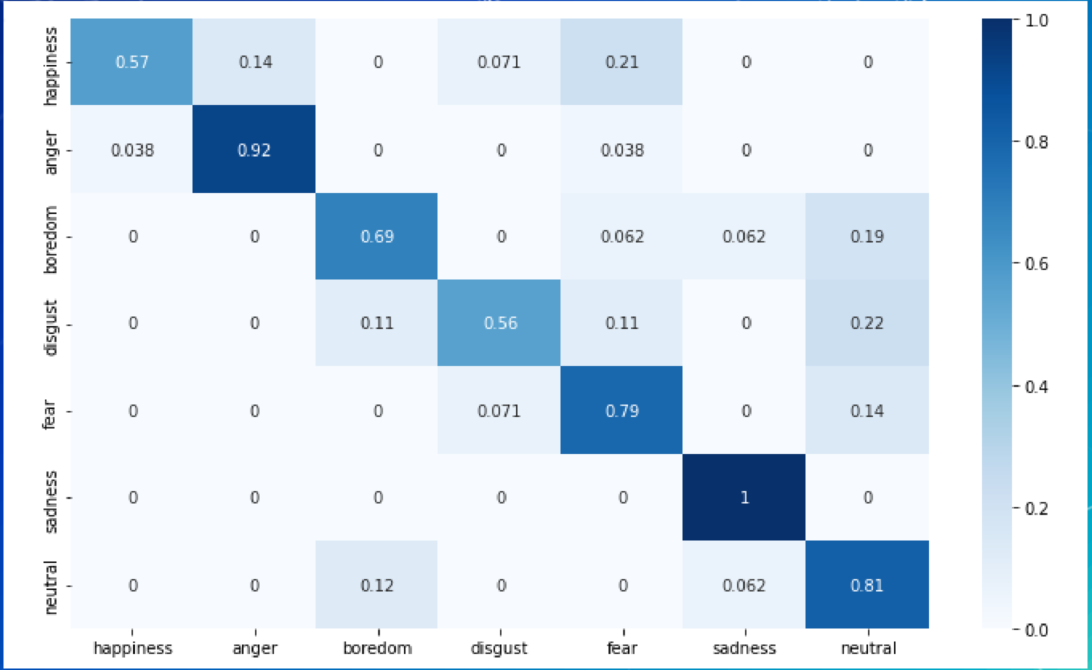


Confusion Matrix

Comment:

Sad speech is the easiest to be predicted.

Happy speech is easy to be classified into anger and fear speech.



	happiness	anger	boredom	disgust	fear	sadness	neutral
happiness	0.57	0.14	0	0.071	0.21	0	0
anger	0.038	0.92	0	0	0.038	0	0
boredom	0	0	0.69	0	0.062	0.062	0.19
disgust	0	0	0.11	0.56	0.11	0	0.22
fear	0	0	0	0.071	0.79	0	0.14
sadness	0	0	0	0	0	1	0
neutral	0	0	0.12	0	0	0.062	0.81

THANKS!

Future ideas:

- LSTM to get the temporal correlation of the features + CNN classifier
- Regularization techniques: Label smoothing and DropBlock to prevent overfitting