Speech Emotion Recognition

Group: A+++++

Introduction

- **Classifcation & Recognition of speech emotions**
- **Toronto emotional speech set**
  - Audio file - does not have observations or features
  - 7 basic emotions, 2 female adults with different ages
- **Why we do this?**

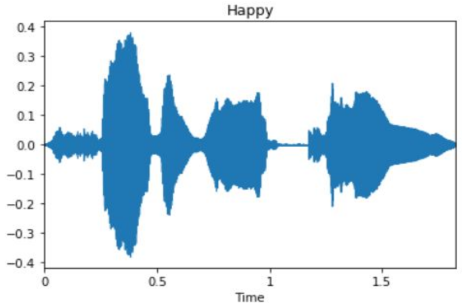


Fig. 2. Wave plot: younger actress say the word "king" with emotion Happy

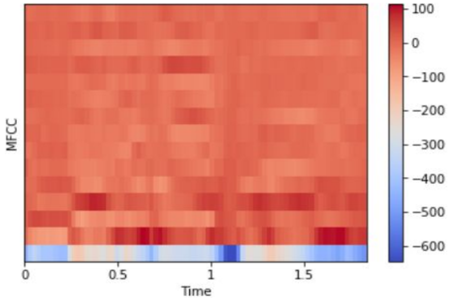


Fig. 3. Data Visualization of MFCC: Happy

Related Work

Psychologist Silvan Tomkins believed there existed a small number (6) of basic emotions that determines the human emotional expression and responses. The aim of emotion recognition is to classify basic emotions, such as fear, happiness, sadness, surprise and anger.

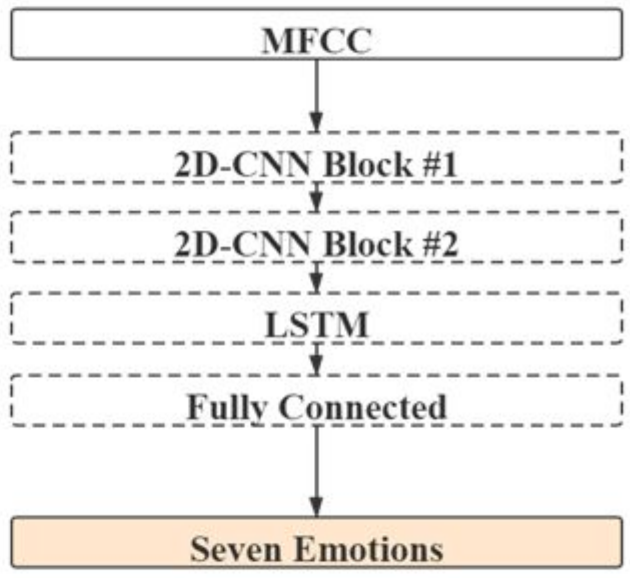
Researchers have explored ways to classify emotions using deep learning in modern day:

1. Examine different acoustic features for emotion recognition in speech using k-means unsupervised clustering. (Iker Luengo & Eva Navas, 2010)
2. Extracting features using Deep Belief Network, then classifying objects using nonlinear SVM classifier. (Chenchen Huang, etc., 2014)
3. Combing the analyses factors, such as facial and bodily expression, heart rate, skin conductivity, thermal signals and brain wave. Analyzing interactions among these factors brings more integrated and context-specific interpretation of expression. (Hatice Gunes $ Maja Pantic, 2010)

This project: takes into account of the continuous characteristics of audio track by combining convolutional neural network and LSTM.

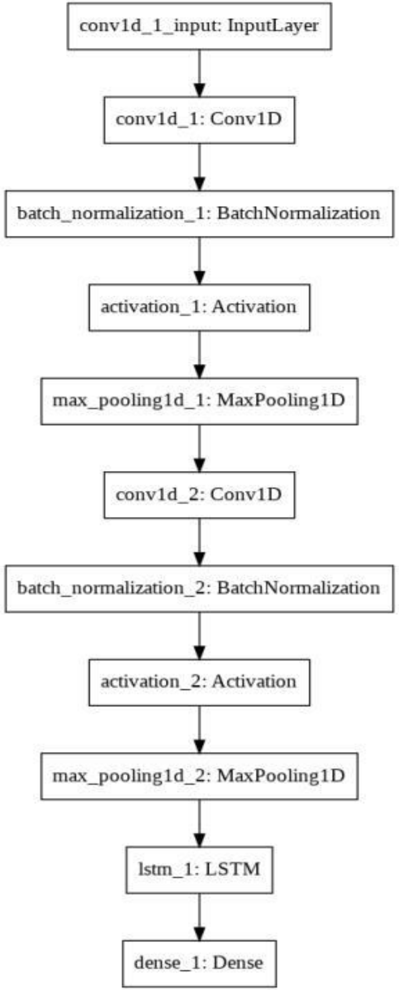
Methods

- Audio file -> Feature Extraction
  - Mel-Frequency Cepstral Coefficient (MFCC)
  - Two form to extract feature to feed into two types of model
  - Train-test split
  - Data preprocessing
- 1D CNN LSTM and 2D CNN LSTM built for comparing
  - Long-term short-term memory method (LSTM)
  - Four parts for both models: two convolutional layers, one LSTM layer and one fully connected layer



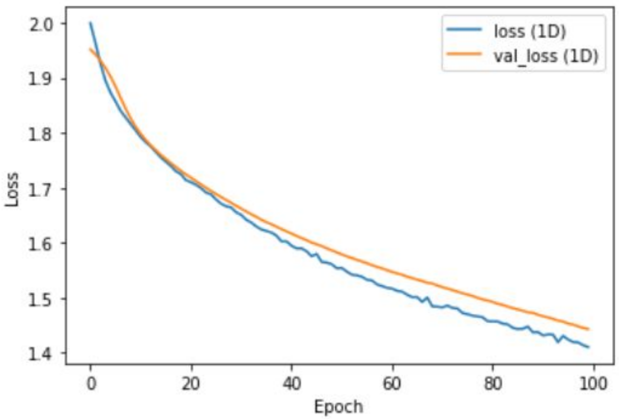
Methods Continued

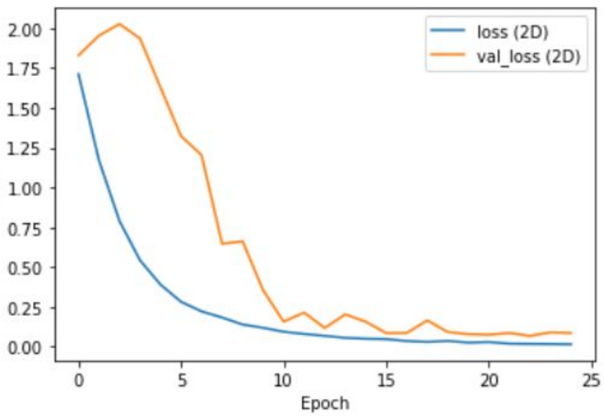
- Each CNN layer contains four parts: one convolution, one ELU activation, one Batch Normalization, and one max pooling
  - Data are reshaped before go through LSTM, and flattened before enter fully connected layer
  - Kernel size of convolution is 3x3, with stride 1x1, pooling size 4x4 and 2x2 with same stride size for pooling layers
  - Softmax activation function is used on the FC layer
- Dropout nodes are added on 2D CNN LSTM for preventing overfitting and results being reviewed

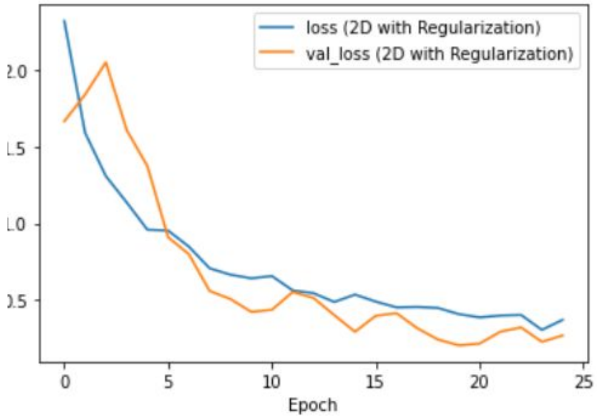

conv1d_1_input: InputLayer
conv1d_1: Conv1D
batch_normalization_1: BatchNormalization
activation_1: Activation
max_pooling1d_1: MaxPooling1D
conv1d_2: Conv1D
batch_normalization_2: BatchNormalization
activation_2: Activation
max_pooling1d_2: MaxPooling1D
lstm_1: LSTM
dense_1: Dense

Result

- 1D-CNN LSTM：low accuracy rate due to processed MFCC bands data
- 2D-CNN LSTM: high accuracy with the risk of overfitting
- 2D-CNN LSTM with Regularization: ensures relatively high accuracy rate while avoiding the problem with overfitting







Conclusion

- Overall Impact
  - Dealing with audio files
  - Feature extraction greatly influence model accuracy
  - RNN
- Limitation
  - Choose between efficiency and accuracy
  - Potential risk of overfitting
- Further Ideas
  - Dataset contain different gender, age, accents
  - Deeper structure