Speech Emotion Recognition Using Convolutional Neural Network and Its Use Case in Digital Healthcare

Nishargo Nigar (ICS)

TUHH
Hamburg University of Technology

RESEARCH OBJECTIVE

- Speech Emotion Recognition (SER) is a field of study that focuses on recognizing human emotions and affective states from speech.
- My project aims to use the Convolutional Neural Network (CNN) to recognize emotions from unseen data (i.e., audio files) and label them according to different emotional ranges using appropriate variables, such as modality, emotion, intensity, repetition, etc.
- Studies have shown that detecting changes in vocal features, such as pitch and speech rate, can provide insight into a patient's emotional state and help diagnose mental health conditions such as depression and anxiety (Girardi et al., 2018).

The research question for the study of this research is:

- **R1**: What is the effectiveness of using Convolutional Neural Networks (CNNs) for detecting different emotions in human speech from unseen audio files, and can this technology be applied to manage depression and anxiety in the field of digital healthcare?

METHODOLOGY

i. Convolutional Neural Network (CNN): A deep learning algorithm utilized for analyzing and processing images.

ii. Short-time Fourier Transform (STFT): The frequency content of a nonstationary signal is examined using the short-time Fourier transform (STFT). The spectrogram time-frequency representation of the signal is defined as the magnitude squared of the STFT.

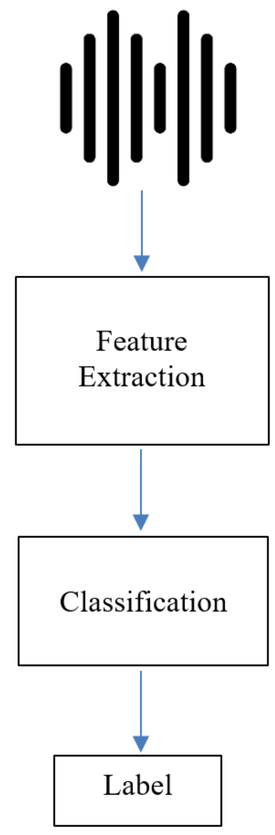

Figure 1: Diagram of the architecture

EXPERIMENTAL SETUP

- Used programming language: Python
- Dataset: RAVDESS

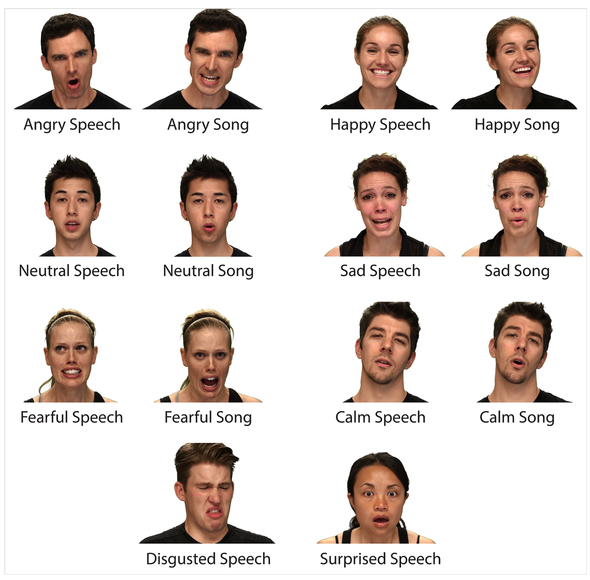

Figure 2: Sample Dataset (Livingstone et al., 2018)

PERFORMANCE EVALUATION

- Training and validation accuracies for CNN
- Loss & Epoch
- Comparison between different models (CNN, LSTM, DNN)

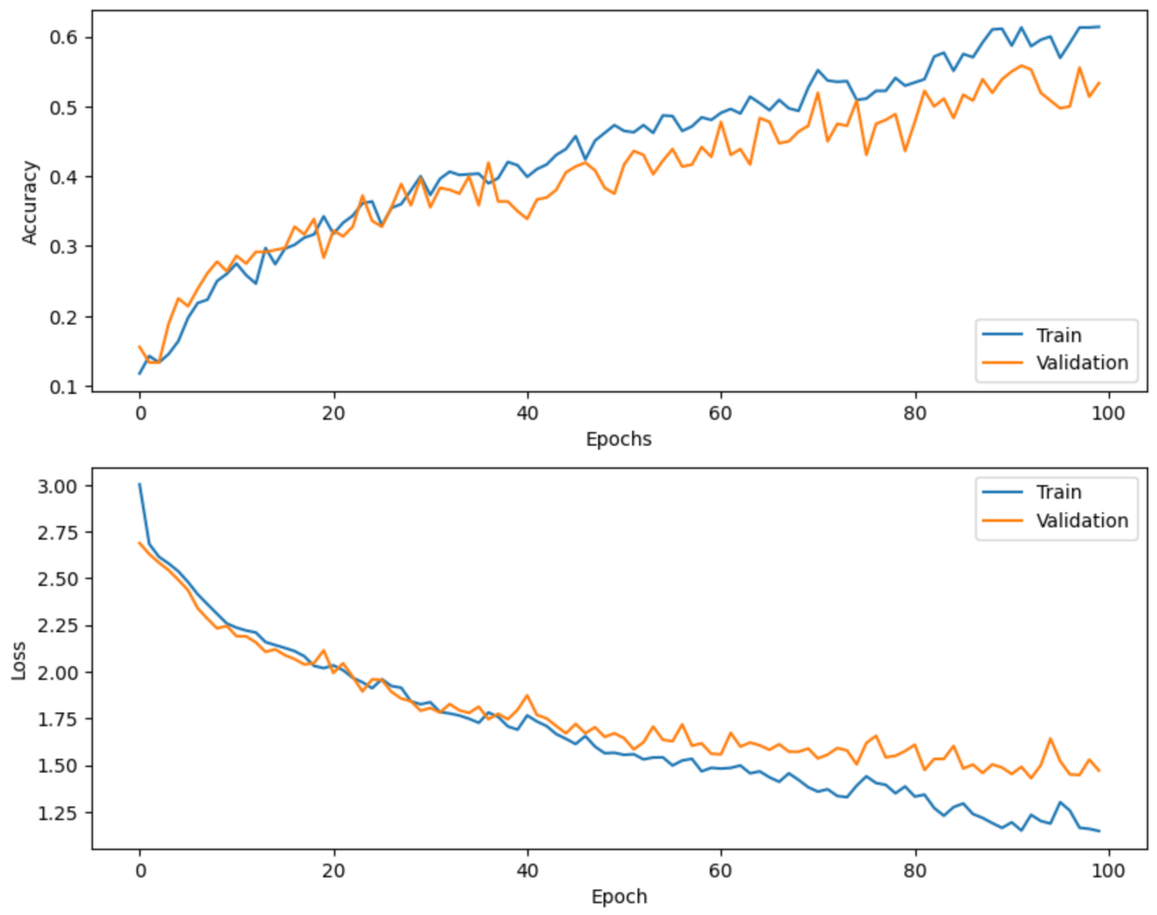

Figure 9: Performance Evaluation for CNN model

Table 1: Comparison Among Different Models

MODEL NAME	PRECISION	RECALL	F1-SCORE	ACCURACY
CNN	0.5444	0.5444	0.5444	0.5444
LSTM	0.4944	0.4944	0.4944	0.4944
DNN	0.5583	0.5583	0.5583	0.5583

EVALUATION: RESULT SUMMARY

- The performance of the DNN model is the best among the three compared models.
- The model for CNN stands in between in terms of performance.
- The precision, recall, F1-Score, and accuracy for the models have a similar pattern. This can happen if the number of False Positives is the same as the number of False Negatives.

CONCLUSION

- I have been able to ascertain that in general, DNN outperforms other models such as CNN and LSTM in this particular case of speech emotion recognition.
- The model creates opportunities in mental health as voice is a predominant factor to recognize the underlying feelings of a human. The use of mobile applications is an example.
- Although the performance of the CNN model was not the best compared to the other models mentioned, there are scopes to iterate and improve.
- Further techniques are required to enhance accuracy and other performance metrics.