Exploring Current Trends in Speech Emotion Recognition: Enhancing Model Effectiveness and Uncovering Multi-Dimensional Use Cases

Bala Chikkala,11588108 Sujit Murahari Giridharan,11589051
Sai Koushik Valluri,11547937 Ratna Srikar Duriseti,11594106
Naga Suchandra Tirumalasetti,11603659

February 2024

1 Introduction

Speech Emotion Recognition(SER) is a technique that is used to automatically detect the human emotions after analyzing the voice signals in the speech. The attributes of the waveform are extracted and depending on the problem statement we filter,normalize,resampling and process the audio files. SER has significance impact in the field of Artificial Intelligence,Natural Language Processing and Large Language Models. SER has gained much popularity in recent years due to its scope of applications in various fields like healthcare, people management, education, etc. Speech emotion recognition can help in identifying people with depression or anxiety issues. It can also assist in analysing the customer sentiment in the field of customer satisfaction analysis.

2 Statement of the Problem

The primary aim of this project is to identify human emotion using a person's speech. Speech Emotion Recognition has gained much popularity in recent

years due to its scope of applications in various fields like healthcare, people management, education, etc

3 Review of Literature

In human and machine communication, SER plays an important role. To contribute towards improved interaction between between humans and machines, recognising emotion from speech can provide useful recommendations [1]. The results showed that the proposed approach works. The model with 1D CNN, LSTM, and attention was more accurate than the model with 2D CNN (88.39 vs. 97.13). It was able to detect multiple emotions from different languages. This was more accurate than previous attempts and showed its generality and effectiveness in SER. The model was more accurate than earlier attempts (88.39) vs. (97.13). The model was also more accurate than previous efforts (68.3 vs. 65.3). The model was able to detect emotions from different datasets such as EMO-DB; SAVEE; ANAD; or BAVED [1].

The author livingstone discourses regarding the emotion speech and song recognition accuracy that is impact its intensity,tempo and voice modulation. And also we took the audio data reference (RAVDESS) from this paper [2].

The article describes a new set of features for automatic emotion recognition from various audio signals. The features are based on perceptual quality metrics and the aim to improve emotion recognition rates that have been compared to existing systems. The study also introduces to a novel set of features that have been focused on the context of speech rather than the conventional features that have been targeted for speech recognition. Also, a majority of the voting decisions rules have been proposed to enhance classifier outputs in the emotion detection tasks. The research tells the emotion classification performance that have been proposed features against the existing systems and which demonstrates the superiority in terms of classification accuracy for valence. Overall, The study aims to advance in the field of emotion recognition in the audio signals which have been an innovative feature for decision making strategies [3].

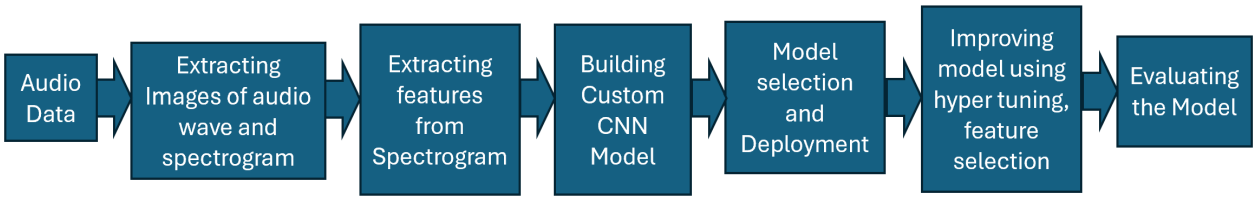
4 Objective of the Study

With the help of deep learning models, we are striving to achieve below provided objectives:
a) Development of custom based CNN models with some assumptions based on the features extracted from the spectrogram-wave forms of the audio files.
b) The finalised CNN model will be saved in a pickle file for the model deployment on Heroku or alternative deployment tools and will be hosted on web using flask API.

5 Research Design and Methodology

5.1 Deep Learning Approaches

We are about to build a deep learning models using three different approaches with different layers. CNN's with different number of hidden layers i.e.,3,4 and 5. And also with epochs of 25,30,45,50, with Learning rate of 0.00001,0.0005,0.000001 will be executed on the audio files. The best out of these parameters will be used to build the final model based on the validation results. The above mentioned parameters are assumptions and we may use different based on the model performance.



6 Data Description

CREMA-D
The data set consists of 7,442 clips from 91 actors, of whom 48 are male and 43 are female, between the ages of 20 - 74, from different varieties of races.

The actors spoke from a selection of 12 sentences, which have six different emotions: anger, disgust, fear, happiness, neutrality, and sadness, as well as four different levels of emotion (low, medium, high, and unspecified).

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)
The dataset consists of 1440 files of 24 professional actors and 60 trials of each actor in different emotions. Out of 24 actors, there are 12 female and 12 male actors vocalizing similar vocabulary that has different meanings in a neutral North American accent. and each emotion includes calm, happy, sad, angry, surprised, and disgusted expressions and is produced at two levels of intensity.

Surrey Audio-Visual Expressed Emotion (SAVEE)
This dataset was developed with the assistance of four native English speakers between the ages of 27 and 31. The total size of the audio dataset is 480. One speaker has 120 occurrences. The emotions that we consider are anger, disgust, fear, happiness, sadness, surprise, and neutral. It consists of six emotions, with each emotion comprising 15 sentences. Within these 15 sentences, there are 3 common sentences for all speakers, 12 distinct emotion-specific sentences, and the remaining 30 sentences are neutral.

Toronto Emotional Speech Set (TESS)
This dataset has been taken from two actresses with a set of 200 target words that were spoken in the phrase "Say the word "and these recordings were made to portray each of the emotions, which include anger, disgust, fear, surprise, sadness, happiness, and neutral. There are around 2800 audio data files in total. It's taken in such a way that the emotions are organized with each of the two female actors and their emotions, which are contained within the folder.

7 Individual Contributions

7.1 Bala Chikkala

Bala is the first person for coming up with the SER idea. Also, he has good amount of knowledge on the process of how the workflow has to be to accomplish the task. He is responsible for collecting the CREMA dataset

that has multiple audio files.

7.2 Sai Koushik Valluri

Defined the methodology to be followed through out the project. Identified different ways in which audio files can be processed and how different deep learning models can be applied to them.

7.3 Sujit Murahari Giridharan

Sujit embarked on compiling the Toronto Speech Dataset after delving into papers and articles. He gathered audio files, ensuring they were arranged based on the emotions conveyed within them. This comprehensive collection serves as a valuable resource for studying speech patterns and emotional expression.

7.4 Ratna Srikar Duriseti

Srikar has come up a few relevant articles and papers. One of the articles that he selected follows the similar approach that we are about to embark. He did collect the SAVEE dataset with the valuable audio files for the project.

7.5 Naga Suchandra Tirumalasetti

Suchandra gathered the RAVDESS audio dataset and proposal document preparation. Currently, he is working on the model development and soon he is going to assist in building the User Interface of our model.

8 Conclusion

This work on Speech Emotion Recognition using custom CNN model architecture to enhance model effectiveness will have a significant implementation in various fields. There will be a continuous improvement in the CNN model architectures that could use attention mechanisms.Finally, a robust model can be built by combining different neural networks and hypertuning the parameters.

References

[1] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern recognition*, 44(3):572–587, 2011.

[2] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

[3] Mehmet Cenk Sezgin, Bilge Gunsel, and Gunes Karabulut Kurt. Perceptual audio features for emotion detection. *EURASIP Journal on Audio, Speech, and Music Processing*, 2012:1–21, 2012.