Grenze International Journal of Engineering and Technology, Jan Issue

Gender Classification using MLP-based Speech Emotion Detection

Kavya S A[1], Kusuma K[2] and Srinivasa K[3]
[1-3]Department of Computer Science and Engineering, Siddaganga Institute of Technology, Tumakuru, India
Email: 1SI19CS059@sit.ac.in, {1SI19CS064, srinivasa}@sit.ac.in

***Abstract*—Speech Emotion Detection (SED) is the extraction of the speaker's emotional state from his or her speech signal. There are a few universal emotions, such as Neutral, Anger, Happiness, and Sadness, that can be trained by any intelligent system with limited computational resources to identify as needed. In this work a gender classification model is proposed using an Multi-Layer Perceptron (MLP) based SED. Because spectral features contain emotional information, they are used in this work. The prosodic features that are used to model different emotions are fundamental frequency, loudness, pitch, speech intensity, and glottal parameters. Each utterance is analysed to extract potential features for the computational mapping of emotions and speech patterns. Classification of Gender is done by detecting pitch using the selected features. Here we used RAVDESS dataset which is trained and tested using a Convolution Neural Network. The most notable features like weight connectivity, local connectivity, and polling results are thoroughly trained and have an accuracy of 52.43 percent. To improve the model accuracy, we used the MLP classifier and feature extraction techniques namely, Mel-Frequency Cepstral Coefficients (MFCC) and Mel spectrogram, which resulted in an accuracy of 68.43 percent.**

***Index Terms*— MLP, Speech detection, Emotion detection, Classification.**

I. INTRODUCTION

Speech is one of the most natural ways for humans to express themselves. We rely on it so much that we recognise its significance when using other forms of communication such as emails and text messages, where we frequently use emoji's to express the emotions associated with the messages. Because emotions play such an important role in communication, detecting and analysing them is critical in today's digital world of remote communication. Because emotions are subjective, detecting them is a difficult task. There is no universal agreement on how to measure or categorise them. Speech emotion recognition is currently an emerging cross-field of artificial intelligence, as well as a hot research topic in signal processing and pattern recognition. The study was widely applied in human-computer interaction [1].

The act of attempting to recognise human emotion and affective states from speech is known as Speech Emotion Detection (SED). This takes advantage of the fact that tone and pitch in the voice frequently reflect underlying emotion. In general, the SED is a computational task that is divided into two parts: feature extraction and emotion machine classification. SED combined with signal processing methodologies may result in Speech-to-Text (STT) technology, which is used as a mode of communication on mobile phones. A SED system is defined as a collection of methodologies for processing and classifying speech signals in order to detect emotions

Grenze ID: 01.GIJET.9.1.540

embedded in them [2].
SED can also be used in a wide range of applications such as interactive voice-based assistant or caller-agent conversation analysis, robots, and banking. It's also found in digital assistants such as Siri, Alexa, Google Now, and Cortana. Another example would be a smart car slowing down when the driver is angry or afraid. As a result, this type of application has a lot of potential in the world and can help businesses.
In this work, we developed a classifier model to classify the geneder of a person using SED which can be used in many applications like crime monitoring by law enforcement agencies.

II. Related Work

Ref. [1] proposed a system for obtaining audio samples of Short-Term Energy (STE), Pitch, and MFCC coefficients in the emotions of frustration, happiness, and sadness. Natural speech was recorded using open source North American English as expression and feedback. As a result, only three emotions were identified: anger, happiness, and sadness. They also identified the speaker's specific characteristics, such as sound, energy, and pitch. In [2] authors proposed a deep learning model of Convolutional Neural Networks (CNN) for SED. Anurag Gupta et.al. [3] proposed a system for resolving the problem of speech emotion recognition by analysing features such as lexical features, visual features, and acoustic features. Ref. [4] considered SED as an exciting ingredient of Human Computer Interaction (HCI). Abhijit Mohanta et.al. [5] have analysed emotions such as angry, fear, happy, neutral from emotional speech signal where features such as loudness, detecting voiced region, excitation energy were used for analysis. Ref. [6] used German Corpus data which had more than 250 recordings. The input data given to Deep Neural Networks (DNN) which has no understanding of the real sense of what the actor is. Michael Neumann et.al. [7] have used t-distributed neighbour embeddings (t-SNE) to analyse visualizations of different representations. Some of the recent works on SED are also identified [8][9][10].

III. Proposed Method

The complete architecture of the proposed SED system is shown in the figure 1, which comprise of mainly three components namely feature extraction from RAVDESS dataset, splitting the dataset into training and testing and Feature classification. The proposed model contains two classifiers CNN and MLP. The one which gives highest accuracy will be selected to display the emotion. Hence, finally emotions are recognized along with accuracy of the overall model.
For a baseline model, we chose a simple classifier that generated predictions by respecting the class distribution of the training data and had a low accuracy score. We then tried a Decision Tree because this is a multi-classification problem that uses average log-mel spectrograms and had a relatively high accuracy score.

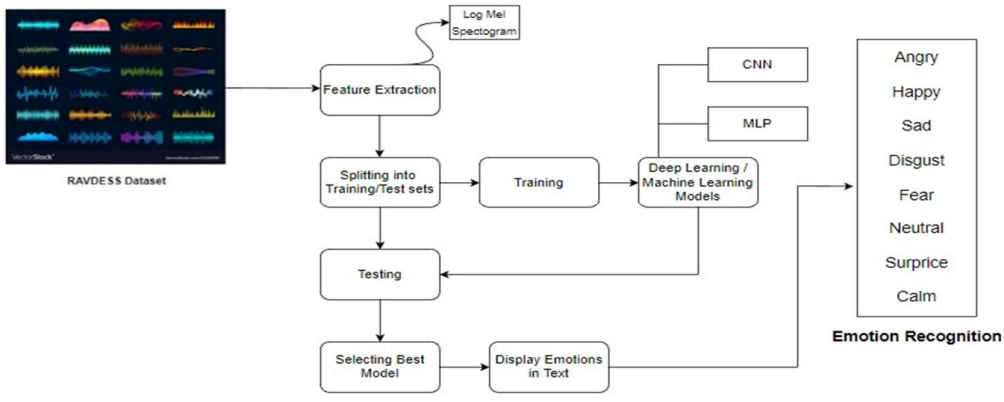

Figure 1. System Architecture of SED

On the test set, we trained a 1D CNN with three convolutional layers and one output layer, which performed slightly better than the baseline decision tree. Locality, weight sharing, and pooling are three key properties of the CNN. Each of them has the potential to improve the performance of speech recognition.
Finally, MLP-Classifier is used to classify the emotions from the given wave signal, which makes the choice of learning rate to be adaptive and yielded maximum accuracy compare to other feature extraction methods.

IV. Result and Comparision Study

Base Model: For a baseline model which uses simple classifier, which generated predictions by respecting the class distribution of the training data and had a low accuracy score of 6.94% shown in figure 2.

```
In [128]: import numpy as np
          from sklearn.dummy import DummyClassifier

          dummy_clf = DummyClassifier(strategy="stratified")
          dummy_clf.fit(X_train, y_train)
          DummyClassifier(strategy='stratified')
          dummy_clf.predict(X_test)
          dummy_clf.score(X_test, y_test)

Out[128]: 0.06944444444444445
```

Figure 2. Accuracy of model using basic Classifier

Feature Extraction using Decision Tree yields an accuracy rate of 32.29%, shown in figure 3.

```
In [129]: from sklearn import tree

          clf = tree.DecisionTreeClassifier()
          clf = clf.fit(X_train, y_train)
          clf.predict(X_test)
          clf.score(X_test, y_test)

Out[129]: 0.3229166666666667
```

Figure 3. Accuracy of model using Decision Tree

Initial Model: 1D CNN with three convolutional layers and one output layer which outputs an accuracy score of 52.43%, shown in figure 4.

```
In [134]: # PRINT LOSS AND ACCURACY PERCENTAGE ON TEST SET
          print("Loss of the model is - " , model.evaluate(X_test,y_test)[0])
          print("Accuracy of the model is - " , model.evaluate(X_test,y_test)[1]*100 , "%")

          9/9 [==============================] - 3s 277ms/step - loss: 1.3371 - accuracy: 0.5243
          Loss of the model is -  1.337149977684021
          9/9 [==============================] - 2s 209ms/step - loss: 1.3371 - accuracy: 0.5243
          Accuracy of the model is -  52.43055820465088 %
```

Figure 4. Accuracy of model using CNN

The accuracy and Loss of the Model is represented graphically in figure 5 and figure 6 respectively. Final Model using MLP Classifier we got an accuracy score of 68.43% as shown in figure 7 and the model is also displaying the emotion associated with each audio files of the dataset.

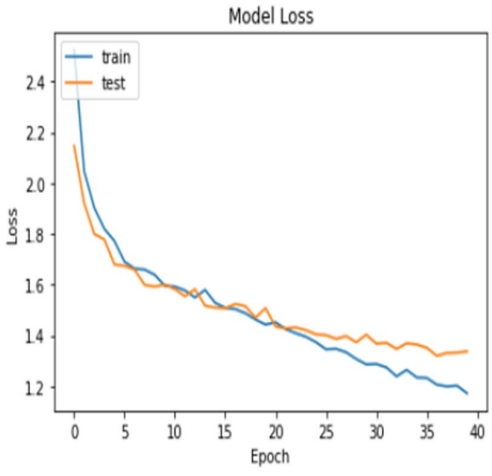

Figure 5. Model accuracy on Train vs. Test dataset

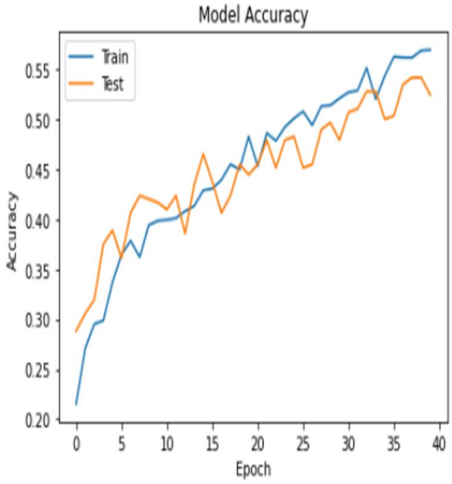

Figure 6. Model loss on Train vs. Test dataset

```
In [67]: while True:
             choice = int(input("Enter 1 to create and train model.  \nEnter 2 to predict on pre-recorded audio. \nEnter 3 to quit. \n"))
             if choice == 1:
                 trainModel()
             elif choice == 2:
                 predictAudio()
             else:
                 quit()

Enter 1 to create and train model.
Enter 2 to predict on pre-recorded audio.
Enter 3 to quit.
1
(576, 192)
Features extracted: 180
Accuracy: 68.23%
Enter 1 to create and train model.
Enter 2 to predict on pre-recorded audio.
Enter 3 to quit.
2
Please enter path to your file.
E:/Speech Emotion Recognition/new/RAVDESS/Actor_01/03-01-01-01-01-01-01.wav
Emotion Predicted: ['happy']
Enter 1 to create and train model.
Enter 2 to predict on pre-recorded audio.
Enter 3 to quit.
2
Please enter path to your file.
E:/Speech Emotion Recognition/new/RAVDESS/Actor_02/03-01-02-01-01-01-02.wav
Emotion Predicted: ['calm']
Enter 1 to create and train model.
Enter 2 to predict on pre-recorded audio.
Enter 3 to quit.
2
Please enter path to your file.
E:/Speech Emotion Recognition/new/RAVDESS/Actor_03/03-01-03-01-01-01-03.wav
Emotion Predicted: ['happy']
```

Figure 7. Accuracy of Model using MLP Classifier

Some of the samples of dataset which are detected based on the actual versus predicted emotions are displayed in the figure 8.

	Actual Values	Predicted Values
140	sad	calm
141	surprise	happy
142	neutral	sad
143	sad	sad
144	fear	surprise
145	sad	sad
146	disgust	disgust
147	angry	angry
148	surprise	surprise
149	angry	angry

Figure 8. Comparision of emotion on Actual vs. Predicted Values

Gender classification based on the predicted emotions is as shown in figure 9.

	gender	emotion	actor	path
0	male	neutral	1	E:/Speech Emotion Recognition/new/RAVDESS/Acto...
1	male	neutral	1	E:/Speech Emotion Recognition/new/RAVDESS/Acto...
2	male	neutral	1	E:/Speech Emotion Recognition/new/RAVDESS/Acto...
3	male	neutral	1	E:/Speech Emotion Recognition/new/RAVDESS/Acto...
4	male	calm	1	E:/Speech Emotion Recognition/new/RAVDESS/Acto...
...	...	...	...	...
1435	female	surprise	24	E:/Speech Emotion Recognition/new/RAVDESS/Acto...
1436	female	surprise	24	E:/Speech Emotion Recognition/new/RAVDESS/Acto...
1437	female	surprise	24	E:/Speech Emotion Recognition/new/RAVDESS/Acto...
1438	female	surprise	24	E:/Speech Emotion Recognition/new/RAVDESS/Acto...
1439	female	surprise	24	E:/Speech Emotion Recognition/new/RAVDESS/Acto...

1440 rows × 4 columns

Figure 9. Classified Genders

V. Conclusion

The accuracy of the Speech emotion recognition system is dependent upon the Database, the features extracted from the databases and a classification model (algorithm) used to classify the emotions. Feature extraction method is discussed for extracting audio features, Log Mel Spectrogram seemed to have an important role in recognizing emotions from speech sample. Choosing the proper and best classifier is very important step in SED and hence, two classifiers being used for classification namely MLP classifier and CNN algorithms are used to recognise the emotion. In this work CNN classifier outputs an accuracy of 52.43% and with MLP classifier we got an accuracy of 68.23%. Finally MLP is used to classify genders from the given voice.

References

[1] Deshmukh, Girija, Apurva Gaonkar, Gauri Golwalkar, and Sukanya Kulkarni. "Speech based emotion recognition using machine learning." In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), pp. 812-817. IEEE, 2019.

[2] M. Wani, T. S. Gunawan, S. A. A. Qadri, H. Mansor, M. Kartiwi and N. Ismail, "Speech Emotion Recognition using Convolution Neural Networks and Deep Stride Convolutional Neural Networks," 2020 6th International Conference on Wireless and Telematics (ICWT), 2020, pp. 1-6, doi: 10.1109/ICWT50448.2020.9243622.

[3] Wadhwa, M., A. Gupta, and P. K. Pandey. "Speech emotion recognition (SER) through machine learning." Analytics Insight (2020).

[4] Khalil, Ruhul Amin, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain. "Speech emotion recognition using deep learning techniques: A review." IEEE Access 7 (2019): 117327-117345.

[5] Ramdinmawii, Esther, Abhijit Mohanta, and Vinay Kumar Mittal. "Emotion recognition from speech signal." In TENCON 2017-2017 IEEE Region 10 Conference, pp. 1562-1567. IEEE, 2017.

[6] Harár, Pavol, Radim Burget, and Malay Kishore Dutta. "Speech emotion recognition with deep learning." In 2017 4th International conference on signal processing and integrated networks (SPIN), pp. 137-140. IEEE, 2017.

[7] Neumann, Michael, and Ngoc Thang Vu. "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech." In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7390-7394. IEEE, 2019.

[8] Huang, Zhengwei, Ming Dong, Qirong Mao, and Yongzhao Zhan. "Speech emotion recognition using CNN." In Proceedings of the 22nd ACM international conference on Multimedia, pp. 801-804. 2014.

[9] Anusha, R., P. Subhashini, Darelli Jyothi, Potturi Harshitha, Janumpally Sushma, and Namsamgari Mukesh. "Speech emotion recognition using machine learning." In 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI), pp. 1608-1612. IEEE, 2021.

[10] Lech, Margaret, Melissa Stolar, Christopher Best, and Robert Bolia. "Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding." Frontiers in Computer Science 2 (2020): 14.