Grenze International Journal of Engineering and Technology, Jan Issue

Speech to Emotion Recognition

P Sai Kameshwari[1], Sai Pragnya[2] and Sairam Utukuru[3]
[1-3]Chaitanya Bharathi Institute of Technology/Information Technology, Hyderabad, India
Email: ugs21094_it.Kameswari@cbit.org.in, ugs21078_it.pragnya@cbit.org.in, usairam_it@cbit.ac.in

***Abstract*— The speech-to-emotion detection deals with detecting emotions through voice. While having a face-to-face conversation with another person, it is often possible to gauge their emotion through cues such as expressions, body language, and more. However, during telephone conversations, it becomes challenging to grasp an individual's emotional state. This work is aimed at recognizing emotions through one's speech, indicating that emotions can be identified by analyzing tone and pitch. This also sheds light on how animals can recognize human emotions.**
In this study, we propose a framework to tackle this challenge. Firstly, we collected a diverse dataset of speech samples annotated with emotion labels. Next, we extracted relevant features from the speech signal, including pitch, energy, MFCCs, and prosodic features. Using this dataset and feature set, we trained machine learning models to classify emotions.

***Index Terms*— Prosodic, Librosa, sklearn, MPL**

I. INTRODUCTION

Emotions play a crucial role in human conversations, often guiding our responses to specific actions or words. A system is a versatile tool in numerous aspects, yet it falls short in capturing human emotions. This work introduces the capability of a machine to discern a person's emotions through voice pitch, intensity, and tone. This ability to recognize emotions holds potential in diverse fields like healthcare, education, and marketing. The primary objective is to categorize emotions into sad, happy, neutral, or angry.

Emotion recognition in speech involves employing machine learning algorithms to analyze the acoustic features of speech, identifying distinctive patterns for different emotions. Real-life applications abound: in customer service, speech emotion recognition can enhance sentiment analysis, facilitating the understanding of customer satisfaction levels and enabling tailored responses. This, in turn, enhances the quality of customer service and fosters loyalty.

In educational settings, speech-emotion recognition can contribute to intelligent tutoring systems. By assessing a student's emotional state during learning activities, the system can adapt feedback and teaching strategies, providing personalized support and motivation. Healthcare also benefits from speech emotion recognition. By analyzing shifts in emotional states over time, it aids in diagnosing and monitoring mental health conditions. Detecting stress, anxiety, or depression indicators from speech signals allows early intervention and appropriate therapies.

Overall, ongoing research and development in speech emotion recognition have the potential to significantly elevate human-computer interactions, positively impacting various aspects of our lives.

Speech emotion recognition (SER) classifies emotions and can even extend to person recognition based on their voice. The subsequent sections of this paper are organized as follows: Section 2 describes related work, Section 3 presents the proposed methodology, Section 4 covers results and discussion, and finally, Section 5 concludes the work.

Grenze ID: 01.GIJET.10.1.501_4

II. Related Work

In accordance with the research paper [1], the work is categorized into two primary steps: feature extraction and emotion classification. During feature extraction, the authors employed MFCCs to extract acoustic features from speech signals. They combined the DNN (deep neural networks) and ELM (Elaboration Likelihood Model theory) to jointly classify emotions. The DNN is responsible for extracting high-level features from the input features, while the ELM is utilized for finalizing emotion classification. The research paper [2] involves employing a DNN model to extract high-level features from speech signals and subsequently aggregating multiple DNN models to construct an ensemble model to enhance classification accuracy. The proposed methodology was evaluated using two datasets: the Berlin Emotional Speech Database and the EmoReact Database. Experimental results indicated that the proposed approach achieved superior classification accuracy compared to various state-of-the-art techniques, including support vector machine (SVM) and k-nearest neighbor (KNN) classifiers.

The research paper [3] introduces an approach that encompasses extracting both spectral and prosodic features from speech signals and fusing them using a feature fusion technique. These fused features are then used as input for an MKL-based classifier, which combines multiple kernel functions to attain improved classification accuracy. In the research paper [4], the process involves training a model on a source domain with ample data and transferring the acquired knowledge to a target domain with limited data. The authors also provide insights into the efficacy of transfer learning across distinct emotional categories and levels of data scarcity.

The authors conduct experiments on two benchmark datasets: the Berlin Database of Emotional Speech and the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset. The research paper [5] entails conducting experiments on the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset, comparing the performance of the proposed multi-model ensemble learning approach to several baseline models, including single SVM, single RF, and single CNN models. In the context of the [6] paper, a hybrid model is proposed that combines a convolutional neural network (CNN) with a long short-term memory network (LSTM) for speech emotion recognition. The feature extraction phase utilizes CNN, while the classification phase employs the LSTM model.

The research paper [7] encompasses three stages: feature extraction, feature fusion, and classification. During feature extraction, both acoustic and text-based features are extracted from speech signals and their corresponding transcripts. Feature fusion involves combining these extracted features using a weighted average, and classification employs a support vector machine (SVM) to categorize speech into distinct emotional categories. The authors discuss the limitations of traditional approaches dependent on acoustic features and conduct experiments on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset.

The approach proposed in research [8] comprises a speech-based model and a text-based model. The speech-based model employs a convolutional neural network (CNN) and a long short-term memory (LSTM) network to extract features from the speech signal, while the text-based model employs a bidirectional LSTM network to process text data. The extracted features from both models are combined using a fusion layer and input into a fully connected network for classification. The authors conduct experiments on the IEMOCAP dataset and compare the performance of their approach to several baseline models.

In the context of the [9] research paper, the same models are used, but five different classifiers are applied, including support vector machines (SVMs), k-nearest neighbor (k-NN), decision tree, naive Bayes, and random forest. The outputs of these classifiers are then combined using an ensemble technique to enhance system performance. The paper [10] also follows the same model structure. The proposed DNN architecture in [11] consists of multiple hidden layers using a rectified linear unit (ReLU) activation function and a SoftMax output layer for emotion classification. The authors also integrate dropout regularization to mitigate overfitting and experiment with various hyper parameters to optimize system performance.

In research paper [11], the authors employed the Emo-DB database for training and testing their system. They extracted both acoustic and linguistic features from speech signals, including pitch, energy, spectral centroid, spectral flux, and occurrences of specific linguistic cues related to emotions. They employed a combination of support vector machines (SVMs) and random forest classifiers for emotion classification.

The research paper [12], following a similar model, compares the performance of their proposed system with other emotion recognition systems such as support vector machines (SVMs) and multi-layer perceptrons (MLPs), all utilizing the same dataset. Experimental results show that their proposed system achieves a superior accuracy of 78.7% on average for the classification of eight basic emotions.

In research [13], the hybrid feature set encompasses both low-level and high-level features extracted from speech signals, including prosodic features, spectral features, and speech signal parameters. The authors employed the Berlin Emotional Speech Database (Emo-DB) dataset to train and test their model. They also compared their

proposed system's performance with other emotion recognition systems using the same dataset. The experimental results demonstrate that their proposed system attains a higher average accuracy of 69.54% for classification.
As per the findings of [14], the authors propose a speech-emotion recognition system using a neural network approach. The authors utilize the MFCC and Perceptual Linear Predictive (PLP) features to extract relevant features from the speech signal. They employ the RAVDESS dataset for training and testing their model. Experimental results demonstrate an accuracy of 63.93% for classification.
In the context of the paper [15], the authors propose a speech emotion recognition system using deep neural networks (DNN) with pre-trained word embeddings. They utilize the EmoReact dataset containing both speech and text data for training and testing. The authors initially leverage a pre-trained word embedding model to convert text data into vector representations. They then employ a DNN to integrate text and speech features for emotion recognition. Experimental results exhibit an accuracy of 72.8% for emotion classification.
In our prior work, we introduced a methodology for analyzing interview performance through facial emotion recognition and speech fluency recognition [16]. Another study addresses the detection of human activity in scenarios involving missing data, specifically for a medium level of missing values.

III. Methodology

In our methodology, we focused on recognizing emotions from speech using an MLP Classifier. We utilized the RAVDESS dataset, which contained 1440 audio files from 24 actors, covering a wide range of emotional expressions, including both strong and weak emotions. The dataset served as the foundation for our research, allowing us to train and evaluate our models effectively. To extract informative features from the audio files, we employed various techniques. We used the sound file library to read the audio files, while the librosa library was utilized to extract essential features such as Chroma and spectrogram. Additionally, we applied normalization and pre-emphasis techniques to enhance the quality of the extracted features. The MLP classifier was chosen as our machine learning model due to its effectiveness in capturing complex patterns and relationships in data. We trained the MLP classifier using the extracted features and the labeled emotional data from the RAVDESS dataset. To evaluate the performance of our model, we created a confusion matrix, which provided insights into the accuracy of emotion recognition. We load the audio files for testing and the emotions present in them are displayed. The modules that are used in this work are: Librosa, NumPy, Pandas, globs, glob, Sound file. As depicted in Figure 1, initially, we need to download the RAVDESS dataset and then commence the program by importing modules such as librosa, NumPy, pandas, sound file, etc. The audio files are available within the RAVDESS dataset; therefore, we proceed to extract features from it using methods like chroma and spectrogram. The spectrogram is defined as a visualization that represents the intensity of the audio. After feature extraction, our next step involves designing a model known as the MLP classifier. This MLP classifier will be categorized into two types: testing and training. The classifier utilized in this study is the MLP classifier.
Figure 1 is the system design. To get started, download the Ravdess dataset and import necessary modules like librosa, NumPy, pandas, and sound file. Extract audio features using chroma and spectrogram methods. Spectrogram reveals audio strength. Utilize an MLP classifier to split the data into training (90%) and testing (10%). This classifier simplifies the process. Compare the accuracy of the two models, with features extracted from the audio files.
Multiple classifiers can be applied in this context, but the MLP classifier simplifies the process by dividing the dataset into training and testing sets. The testing set comprises 10% of the dataset, while the remaining 90% is allocated for training. The training model is more efficient compared to the testing model, given that it utilizes 90% of the dataset for training purposes.
Initially, the dataset contains audio files. The subsequent step involves feature extraction from these files using chroma and spectrogram techniques. The output of the feature extraction serves as input for the MLP classifier. Subsequently, the MLP classifier is divided into training and testing models, allowing for a comparison of their respective accuracies.

IV. Results And Discussion

Firstly, in this work we need to import the required modulus. Figure 2 represents the wave plot concerning time. Next, we need to create a wave plot for the voices present in the ravdess dataset. wave plot is a graph that is drawn between emotions and time is drawn between emotions and time.

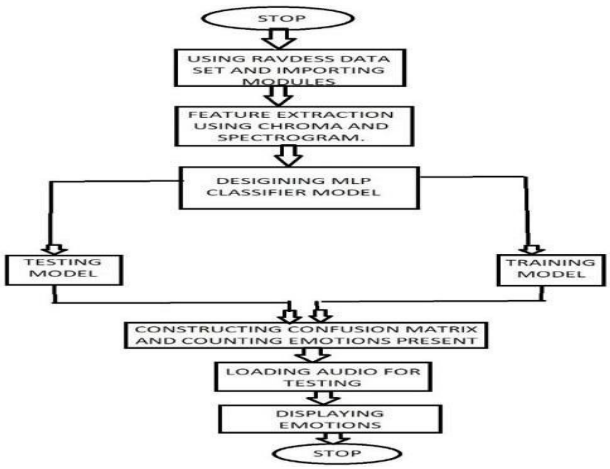

Figure 1: System design

```
librosa.display.waveplot(x,sr=sr)
```

```
<matplotlib.collections.PolyCollection at 0x7f98c965c310>
```

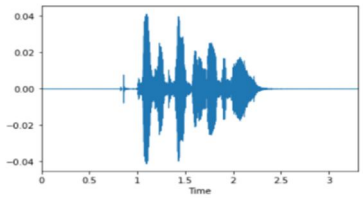

Figure No: 2 wave plot concerning time

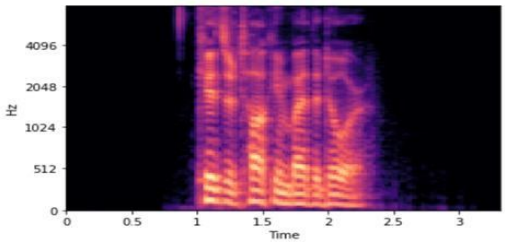

Figure No: 3 Spectrogram

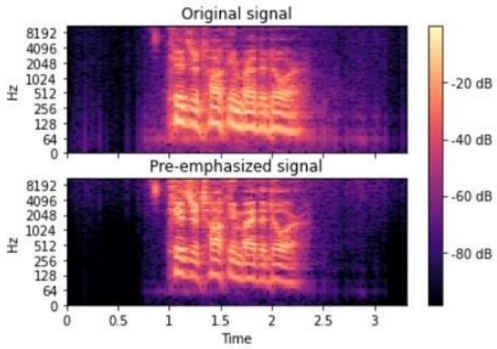

Figure No:4 speech signal

Figure 3 represents Spectrogram, which represents how strong or weak the emotions are. A spectrogram is used for feature extraction. A spectrogram is used to represent the strength of the signal at different frequencies.
Figure 4 represents the speech signal. The speech signal contains many parts of silence. These silent parts are not required for us, and this should be removed for accurate output. If we do not remove these silence parts, then the

emotions displayed may be incorrect so removal of these silence parts in the speech signal is essential and the graphs between frequency and time before and after removal of the speech signal are similar.
Figure 5 represents the pre-emphasis of the speech signal which is the most important step of pre-processing at high frequency. It finds comparable amplitude by passing the speech signal through a high-pass filter (FIR). Now, we need to extract features from audio files using the chroma method. We need to load all the audio that is present in the files. The classifier used in this work is MLP. MLP classifier makes the work easy by splitting the dataset into test and train models. 10% of the dataset is used for testing and the remaining 90% is used for training. The training model is more efficient than the testing model as 90% of the dataset is used for training. Initially, the audio files are present in the dataset. Then the next step is to extract the features from it using chroma and spectrogram. The output of the feature extraction is given as input to the MLP classifier. Then the MLP classifier gets split into test and train models and their accuracy can also be compared with each other classifier will classify different emotions such as anger disgust, sad, etc.
Figure 6 represents the graph that classifies the emotions percentage. Now we need to construct the confusion matrix and count the number of emotions present in the audio file. Figure 7 is loading the audio for testing. The Audio is loaded in test audio and the location of the audio that needed to be loaded for classifying the emotions is stored.

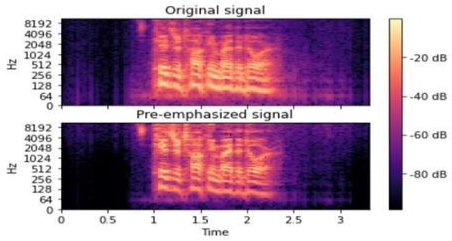

Figure No: 5 pre-emphasis of the speech signal

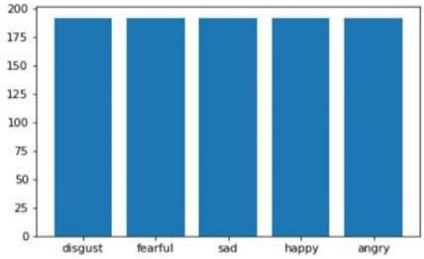

Figure No 6: emotions percentage

Loading a audio for testing

```
In [35]:
test_audio, sr = librosa.load('../inpu
t/ravdess-emotional-speech-audio/Actor
_04/03-01-04-01-02-02-04.wav')
Audio(data=test_audio, rate=sr)
```

Out[35]:

-0:03

Figure No: 7 loading the audio for testing

```
# Actual Emotion
emotion
```

Out[37]:

'sad'

Figure No: 8 displays the final emotion

Figure 8 displays the final emotion. Hence depending upon the emotion whether it is sad or angry or happy etc. can be displayed in the output.

V. Conclusion

In conclusion, our methodology allowed us to recognize emotions from speech using an MLP Classifier. We utilized the RAVDESS dataset, which provided a diverse range of emotional expressions for training and evaluation. By employing techniques such as feature extraction using Chroma and spectrogram, and training our MLP classifier on the extracted features, we were able to effectively recognize emotions from speech signals.
Moving forward, there is scope for further development in speech emotion recognition. Future work involves building a more advanced Speech Emotion Recognition (SER) model that can handle emotions from multiple speakers and accurately capture emotions even when the speaker speaks quickly. This would enhance the model's applicability and broaden its potential real-world use cases.

References

[1] Nguyen, T. H., Doan, D. H., & Nguyen, T. N. (2021). Speech emotion recognition using deep neural networks and extreme learning machine. Journal of Ambient Intelligence and Humanized Computing, 12(3), 2683-2696. DOI: 10.1007/s12652-020-02633-2.

[2] Shang, R., Zhang, Q., & Zhang, X. (2021). Speech emotion recognition based on deep neural network and ensemble learning. Journal of Computational Science, 54, 101364. DOI: 10.1016/j.jocs.2021.101364.

[3] Wang, X., Wang, Y., & Zhu, Q. (2019). Speech emotion recognition with feature fusion and multiple kernel learning. IEEE Transactions on Affective Computing, 12(3), 508-518. DOI: 10.1109/TAFFC.2019.2894661.

[4] Nguyen, H. T., Nguyen, V. T., & Tran, T. N. (2019). Exploring the effectiveness of transfer learning for speech emotion recognition. IEEE Access, 7, 171275-171285. DOI: 10.1109/ACCESS.2019.2953166.

[5] Zhu, S., Liu, J., & Liu, Q. (2021). Speech emotion recognition with multi-model ensemble learning. IEEE Access, 9, 107142-107152. DOI: 10.1109/ACCESS.2021.3109401.

[6] Sankarasubramaniam, Y., & Soman, K. P. (2020). Speech emotion recognition using convolutional neural network and long short-term memory network. International Journal of Advanced Computer Science and Applications, 11(11), 434-440.

[7] Chowdhury, A., & Narayanan, S. (2020). Speech emotion recognition using fusion of acoustic and text features. IEEE Transactions on Affective Computing, 11(2), 214- 225. DOI: 10.1109/TAFFC.2018.2797862.

[8] Narayan-Chen, A., Shah, Y., & Bellamy, R. K. E. (2020, June). Multimodal emotion recognition from speech and text using deep learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp.2879-2888). DOI: 10.1109/CVPRW50498.2020.00359.

[9] Kumar, A., Jain, S., & Kumar, D. (2015). Emotion recognition from speech using prosodic and spectral features with ensemble of classifiers. In Proceedings of the International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 655- 661).

[10] Shukla, M., Rao, K. S., & Uma Maheshwari, S. (2016, September). Emotion recognition from speech using deep learning techniques. In Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA) (pp. 225- 232).

[11] Goel, N., & Tiwary, U. S. (2018, November). Emotion detection from speech using multi-modal techniques. In Proceedings of the International Conference on Intelligent Systems Design and Applications (ISDA) (pp. 374-379).

[12] Siddiqui, S. A., & Jha, R. K. (2019, December). Emotion recognition from speech using convolutional neural network with transfer learning. In Proceedings of the International Conference on Computing and Communication Systems (I3CS) (pp. 468-475).

[13] Verma, H., Dubey, A. K., & Singh, M. K. (2019, February). Speech emotion recognition using hybrid features and stacking ensemble classifier. In Proceedings of the International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1033- 1040). IEEE.

[14] Kaushik, H., Kumar, S., & Joshi, N. (2017). Emotion recognition from speech using neural network approach. In Proceedings of the International Conference on Computational Intelligence and Communication Networks (CICN) (pp. 363-368). IEEE.

[15] Panda, S., & Bansal, P. (2019, July). Speech emotion recognition using deep neural networks with pre-trained word embeddings. In Proceedings of the 3rd International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 1132-1136). IEEE.

[16] Y. Adepu, V. R. Boga and S. U, "Interviewee Performance Analyzer Using Facial Emotion Recognition and Speech Fluency Recognition," 2020 IEEE International Conference for Innovation in Technology (INOCON), Bengaluru, India, 2020, pp. 1-5, Doi: 10.1109/INOCON50539.2020.9298427.