Detection of Speech Emotions Using Deep Learning Techniques

Aswin Manoj
Dept of Artificial Intelligence & Data Science
Paavai College of Engineering
Namakkal , India
aswinmanoj0306@gmail.com

Arjun Sisupal
Dept of Artificial Intelligence & Data Science
Paavai College of Engineering
Namakkal , India
arjunsisupal18@gmail.com

Sooryan A.P
Dept of Artificial Inteeligence & Data Science
Paavai College of Engineering
Namakkal , India
sooryanap2002@gmail.com

Raguraman P.J (A.P AI&DS)
Dept of Artificial Intelligence & Data Science
Paavai College of Engineering
Namakkal , India
raguramanjayagopalpce@paavai.edu.in

Jayan Babu M
Dept of Artificial Intelligence & Data Science
Paavai College of Engineering
Namakkal , India
jayanbabu6143@gmail.com

***Abstract*—The need for recognizing different emotions through speech is increasing day by day. Speech Emotion Recognition (SER) is the endeavor to discern human affective and emotional states from speech. Numerous strategies have been used to extract emotions from signals, including various well-established feature extraction and classification techniques of which Deep Learning techniques are gaining prominence. Many studies show that mental healthcare often requires gender-specific treatments. The existing SER systems do not incorporate gender information about the speaker and hence cannot be applied in telemedicine, especially for mental health care services**

where gender-specific treatments are necessary. This paper focuses on the relevance of gender information incorporated into SER systems and their applications in mental health care services.

In this paper, Convolutional Neural Network (CNN) and CNN with Long Short-Term Memory

(LSTM) are the models used and the results are compared for choosing the best model. From the results, it is observed that SER using CNN-LSTM has high accuracy compared to the CNN model. Gender information is incorporated in both CNN and LSTM models to check if it contributes to the increase or decrease in the accuracy of the models. The dataset used is RAVDESS. A web application for SER using CNN-LSTM is implemented considering six emotion categories. This

paper also includes a detailed review of the dataset, feature extraction methods, augmentation of the dataset, preparation of the dataset, and graphs associated with it.

***Keywords*—speech emotion detection, deep learning techniques, gender specific treatments, convolutional network**

1. INTRODUCTION

Speech Emotion Recognition (SER) has evolved from a specialized field to a pivotal elementof Human-Computer Interaction (HCI). These systems leverage direct speech interaction instead of conventional input devices to comprehend verbal information and facilitate responses from human users [1]. Notable applications encompass voice emotion-based medical

solutions, in-car navigation systems, and conversation systems for call centers. However, HCI systems encounter significant challenges during the transition from laboratory testing to real world implementation. Consequently, there is a need for proactive initiatives aimed at effectively resolving these challenges and advancing machine emotion recognition.Various machine learning models, both linear and non-linear, can be employed for Speech Emotion Recognition (SER). Linear models include Support Vector Machines (SVM),Bayesian Networks (BN), and Maximum Likelihood Principle (MLP). However, due to the non-stationary nature of speech signals [1], non-linear classifiers such as Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), Gaussian Mixture Model (GMM), and Hidden Markov Model (HMM) are believed to offer advantages for SER tasks.

Deep Learning has garnered increased attention and has become a prominent research domain within Artificial Intelligence in recent years. In the context of Speech Emotion Recognition (SER), deep learning techniques offer notable advantages over conventional methods. These include the ability to automatically detect complex structures and features without the need for manual feature extraction and tuning, the capability to extract low-level features directly from raw data, and the aptitude to handle unlabeled data. Deep learning techniques utilized for SER encompass Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), Recurrent Neural Networks (RNN), and Deep Neural Networks (DNN), among others. CNNs excel in learning features from high-dimensional input data, although they are susceptible to capturing minor fluctuations and distortions, necessitating significant storage capacity.

Deep learning techniques have superseded traditional approaches due to their inherent characteristics such as feature learning, end-to-end learning, representation learning, capacity to handle complex relationships, scalability, transfer learning, and adaptability. The CNN-LSTM model utilized in speech emotion recognition effectively integrates the capabilities of CNNs for learning high-level image features with the strengths of LSTM networks in capturing temporal dependencies within sequential data. Empirical evidence showcases the state-of-the-art performance achieved by this model in speech emotion recognition tasks.

II. LITERATURE SURVEY

In [4], a unique emotion recognition method that does not rely on any speech acoustic data and incorporates speaker gender information was suggested. This seeks to gain from the rich information from speech raw data, without any artificial intervention. In general, voice emotion identification systems require user selection of acceptable traditional acoustic parameters as classifier input for emotion recognition. Utilizing deep learning methods, the network automatically selects important information from raw voice signals for the classification layer to complete emotion recognition. It can prevent the omission of emotional information that cannot be directly analytically described as a speech acoustic characteristic. Adding speaker gender information to the suggested method greatly increased recognition accuracy. The proposed approach in [4] combines a Residual Convolutional Neural Network (R-CNN) and a gender information block. The raw voice data is supplied to these two blocks concurrently. The R-CNN network gathers the necessary emotional information from the voice input and classifies the emotional category. The suggested technique is examined on three public datasets with distinct language systems. Experimental results reveal that the suggested method offers 5.6%, 7.3%, and 1.5%, respectively accuracy gains in Mandarin, English, and German compared with existing highest-accuracy algorithms. In order to verify the generalization of the proposed approach, FAU and eNTERFACE databases are employed and achieved 85.8% and 71.1% accuracy, respectively.

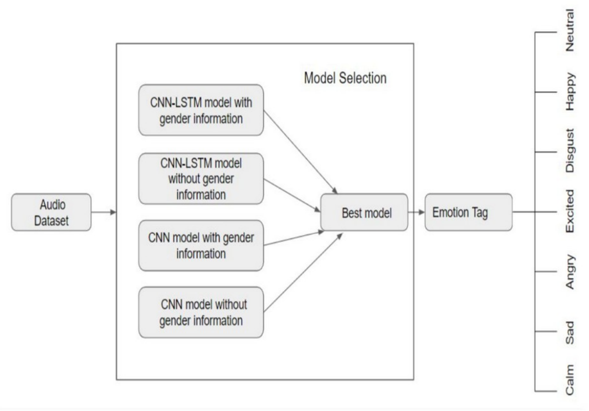
Deep Neural Networks (DNNs) denote multilayer artificial neural networks with more than one hidden layer and millions of free parameters. In [5], a Generalized Discriminant Analysis (GerDA) based on DNNs to learn discriminative features of low dimension optimized with respect to a quick classification from a wide range of acoustic characteristics for emotion recognition was presented.On nine regularly used emotional speech corpora, the performance of GerDA features and their subsequent linear classification with previously reported benchmarks produced using the same collection of acoustic features identified by Support Vector Machines (SVMs) was compared in [5]. The results impressively indicated that low-dimensional GerDA features capture hidden information from the acoustic characteristics leading to a dramatically raised unweighted average recall and a greatly raised weighted average recall.

Automatic emotion recognition from the speech is a tough problem that relies greatly on the efficacy of the speech elements employed for classification. In the work done in [6], the application of deep learning to automatically discover emotionally important elements from the speech was examined. It was shown that utilizing a deep recurrent neural network, both the short-time frame-level acoustic aspects that are emotionally significant, as well as a suitable temporal aggregation of those elements into a compact utterance-level representation was learned. Moreover, a novel technique for feature pooling over time which employs local attention in order to focus on specific portions of a speech signal that is more emotionally salient was proposed in [6]. The suggested approach was assessed on the IEMOCAP corpus and was demonstrated to produce more accurate predictions compared to existing emotion identification systems.

III. PROPOSED SYSTEM

The audio dataset used for the proposed SER system is RAVDESS. The dataset is beinginputted into the four models: CNN with gender information, CNN without gender information, CNN-LSTM with gender information, and CNN-LSTM without gender

information. After training each model with a certain number of epochs, the accuracy is measured. The model with the highest accuracy is chosen as the best model. The output of the model is an emotion tag which is predicted from the inputted audio. Emotion tags are of the following categories: Neutral, Happy, Disgust, Excited, Anger, Calm, Sad.


Audio Dataset
Model Selection
CNN-LSTM model with gender information
CNN-LSTM model without gender information
CNN model with gender information
CNN model without gender information
Best model
Emotion Tag
Neutral
Happy
Disgust
Excited
Angry
Sad
Calm

IV. PROPOSED ALGORITHM

The difficult challenge of analyzing and interpreting human speech to ascertain the underlying emotion is known as speech-emotion recognition. To precisely recognize and categorize the emotions contained in spoken language, the speech emotion identification model algorithm was created. This model architecture, which is based on deep learning, processes the input and extracts pertinent information for classification using multiple layers of neural networks. Preprocessing the voice signal is the initial stage in the speech emotion recognition model method. Mel Frequency Cepstral Coefficients (MFCCs), a method, are used to convert the voice signal into a series of characteristics. This method is frequently applied in speech processing applications because it isolates the most important information from the speech signal, such as the frequency bands and their relative strengths. The neural network model receives the voice signal after it has been converted into a series of MFCCs. Each layer of neural network in the model architecture serves a particular purpose in the classification process. The input is processed by a Conv1D layer in the first layer to identify pertinent features. In order to inject non-linearity into the model, the output of this layer is routed through a Rectified Linear Unit (ReLU) activation function.14 A MaxPooling1D layer follows, which down samples by taking the highest value found within a pool window. This layer reduces the input's spatial size, which in turn lowers the model's computing burden and parameter count. Conv1D and MaxPooling1D layers with progressively more filters are added to the model in order to execute more intricate feature extraction. To enhance the stability and efficiency of the model, batch normalization and dropout regularization are also applied to the output of some of these layers. A fully connected layer with output units equal to the number of classes in the task of classifying emotions makes up the model's top layer. To determine class probabilities, the output of this layer is subjected to a softmax activation function. Backpropagation is a method used to train the voice emotion recognition model algorithm. During training, the neural network model's weights and biases are updated using the backpropagation technique. A dataset of voice sounds with accompanying emotion labels is used to train the model. Categorical cross-entropy, a measure of the discrepancy between the predicted class probabilities and the actual class labels, is the loss function utilized in training. Once trained, the model can be used to anticipate the emotions expressed in fresh voice signals. The trained neural network model is then fed the preprocessed speech signal using the MFCC approach for classification. The projected class probabilities for each of the potential emotions are output by the model. In conclusion, the speech emotion identification model algorithm uses deep learning to precisely recognize and categorize the emotions that are expressed in spoken language. The approach involves employing MFCCs to preprocess the speech signal before running it through a number of layers of neural networks to extract and classify features. The model, which may be used to forecast emotions in fresh voice data, was trained using backpropagation and categorical cross-entropy loss.

Two CNN models are implemented for SER, one is using gender information, and the other, without using gender information. The CNN model using gender information takes gender as 15 a feature from the audio dataset, and hence an extra parameter is taken into consideration which is expected to improve the accuracy of the model.

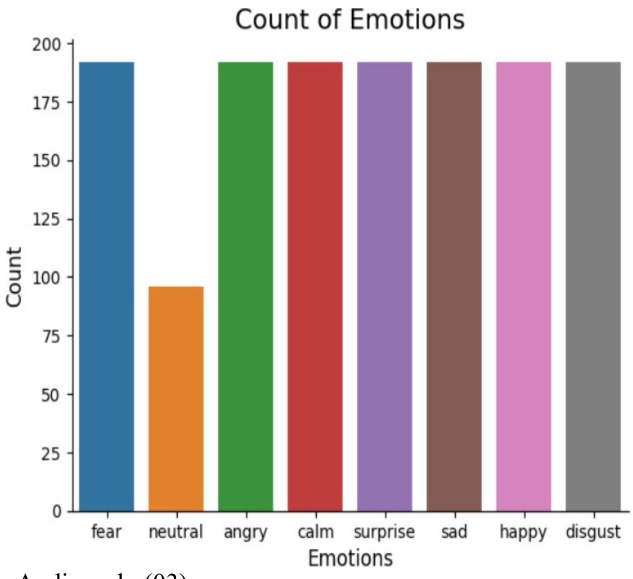
The CNN model without using gender information does not take gender as a feature from the audio dataset and hence the accuracy compared to the previous model is expected to be less.

V. RESULT AND DISCUSSION

The filename consists of a 7-part numerical identifier (e.g.,03-01-06-01-02-01-12.wav). These identifiers define the stimulus characteristics: Modality (01 = full-AV, 02 = video-only, 03 = audio-only), Vocal channel (01 = speech, 02 = song), Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised), Emotional intensity (01 = normal, 02 = strong). NOTE: There is no strong intensity for the 'neutral' emotion, Statement (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door"), Repetition (01 = 1st repetition, 02 = 2nd repetition), Actor (01 to 24. Odd numbered actors are male, even-numbered actors are female).

Dataset Name	Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [1]
Actors	24 Actors (12 Male and 12 Female)
Emotions	Happy, sad, angry, surprised, disgust, fear, and neutral expressions
Number of files	7356
Statements	“Kids are talking by the door.”, ”Dogs are sitting by the door.”

Filename example: 03-01-06-01-02-01-12.wav



-Audio-only (03)

-Speech (01)

-Fearful (06)

-Normal intensity (01)

-Statement "dogs" (02)

-1st Repetition (01)

-12th Actor (12)

V. CONCLUSION

In this paper, speech emotion is recognized using CNN and CNN-LSTM models with and without using gender information. The models are trained using the RAVDESS dataset which is one of the most popular datasets for audio and video recognition. Based on the evaluation metrics such as accuracy, precision, recall, and F1-score, it was found that the CNN-LSTM model outperformed the CNN model in recognizing emotions accurately. The CNN-LSTM model without using gender information achieved an accuracy of around 70%, compared to the CNN model without using gender information, which achieved an accuracy of around 60%. The CNN-LSTM model using gender information achieved an accuracy of 85%, compared to the CNN model using gender information, which achieved an accuracy of 72%. From the four models mentioned earlier, it was found that CNN-LSTM using gender information has the highest accuracy. It was discovered that combining convolutional and long short-term memory networks (CNN-LSTM) with using gender information resulted in higher performance, demonstrating that temporal information in voice signals was critical in accurately recognizing emotions. The usage of MFCC for feature extraction also helped the CNN-LSTM model perform better.

REFERENCES

[1] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," in IEEE Access,vol. 7, pp. 117327- 117345, 2019, doi:10.1109/ACCESS.2019.2936124.

[2] Fiona Judd, Sue Armstrong & Jayashri Kulkarni (2009) Gender-sensitive mental health care,Australasian Psychiatry, 17:2, 105-111, DOI: 10.1080/10398560802596108

[3] Afifi, Mustafa. (2007). Gender differences in mental health. Singapore medical journal. 48.385-91.

[4] Sun, Ting-Wei. (2020). End-to-End Speech Emotion Recognition With Gender Information. IEEE Access. PP. 1-1. 10.1109/ACCESS.2020.3017462.

[5] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, ''Deep neural networks for acoustic emotion recognition: Raising the benchmarks,'' in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2011, pp. 5688–5691.

[6] S. Mirsamadi, E. Barsoum, and C. Zhang, ''Automatic speech emotion recognition using recurrent neural networks with local attention,'' in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Mar. 2017, pp. 2227–2231.

[7] J. Deng, X. Xu, Z. Zhang, S. Fruhholz, and B. Schuller, ''Semisupervised autoencoders for speech emotion recognition,'' IEEE/ACM Trans. Audio, Speech, Lang., Process., vol. 26, no. 1, pp. 31–43, Jan. 2018.

[8] V. A. Petrushin, ''Emotion recognition in speech signal: Experimental study, development, and

application,'' in Proc. 6th Int. Conf. Spoken Lang. Process., Beijing, China, 2000, pp. 222–225.

[9] Oliveira, Jorge & Praça, Isabel. (2021). On the Usage of Pre-Trained Speech Recognition Deep Layers to Detect Emotions.IEEEAccess.PP.11.10.1109/ACCESS.2021.3051083.

[10] M. Wollmer, B. Schuller, F. Eyben, and G. Rigoll, ''Combining long short term memory and

[11] Jonte Dancker (2022) Brief Introduction to Recurrent Neural Networks: An Introduction to RNN,