2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)
Amity University, Noida, India. Mar 14-15, 2024

2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) | 979-8-3503-5035-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICRITO61523.2024.10522204

Speech Emotion Recognition using Deep Learning

Divya Garg
Computer Science and Engineering Department
KIET Group of Institutions
Ghaziabad, India
divyagarg2001@gmail.com

Amrita Singh
Computer Science and Engineering Department
KIET Group of Institutions
Ghaziabad, India
amritasingh09275@gmail.com

Anshika Gupta
Computer Science and Engineering Department
KIET Group of Institutions
Ghaziabad, India
1anshika.india@gmail.com

Sanjiv Sharma
Computer Science and Engineering Department
KIET Group of Institutions
Ghaziabad, India
martin.mmmec@gmail.com

***Abstract*— Speech Emotion Recognition (SER) is an emerging field that involves recognizing emotions conveyed in speech. Emotions expressed through speech can greatly impact decision-making. This paper delves into the topic of speech emotion recognition (SER) and its focus on interpreting emotions conveyed through spoken language. The importance of SER lies in its potential to improve human-computer interaction, cognitive analysis, and psychiatric assessment. The study combines and preprocesses audio data from various datasets, such as RAVDESS, CREMA-D, TESS, and SAVEE, and uses log mel spectrograms to effectively extract features. Various methods including CNN models, and standard and optimized feature extraction techniques are used. The results suggest that SER has significant real-world applications and the approaches provided effectively identify emotional and voice signals.**

Keywords*— SER, CNN, MFCC, SVM, Speech Emotion Recognition, Deep Learning

I. INTRODUCTION

One of the best and most natural ways for people to talk to each other is through speech signals. Voice signals are generally thought to be the fastest and most useful way for people and computers to talk to each other [1]. Everyone automatically uses all of their senses to get the most out of a message they receive. For machines, detecting feelings is very hard, but for people, it's easy. Machines and people can communicate better with each other when systems know about feelings [3]. Language-based speech emotion recognition can tell the difference between the feelings of male and female speakers. LPCC, MFCC, and Fundamental Frequencies are some of the speech traits that have been studied. People who study speech perception start with these traits. No clear reason is given for why it's hard to tell the difference between different feelings from people's speech [4]. How people talk is affected by differences in their speaking rates, styles, sentence structures, and speakers. Some people might feel different emotions when they hear the same word, which can make it hard to tell the difference between them [5]. Cultural and environmental factors affect how feelings are shown by the person speaking. This makes things more difficult because different cultural and environmental factors lead to different speaking styles. Feelings can be broken down into two groups: short-term and long-term. The recognizer does, however, not say what kind of feeling it found [6]. People can be speaker-independent or speaker-dependent when it comes to recognizing feelings in speech.

KNN, SVM, and Convolutional Neural Network are some classification methods [7]. Talk emotion recognition is briefly explained at the beginning of this work, followed by a block diagram of a voice emotion recognition system. In the next section, we'll talk about some of the datasets that have been the subject of a lot of studies. Analyzing emotional speech and different types of speech falls under this area. When it comes to speech emotion recognition, the fourth section gives short explanations of the different ways that features can be extracted. It then goes into more depth about the classification part. K-nearest neighbor (KNN), support vector machines (SVM), convolutional neural networks (CNN), recurrent neural networks (RNN), and other machine learning algorithms have been talked about in this part. When it comes to voice emotion recognition, the fourth part talks briefly about deep learning.

This paper is motivated by SER's potential to enhance human-computer interaction, cognitive analysis, and psychiatric assessments, this study presents novel contributions in methodology and application. Leveraging diverse datasets like RAVDESS, CREMA-D, TESS, and SAVEE to preprocess audio data and utilise log mel spectrograms for feature extraction. Employing Convolutional Neural Network (CNN) models and optimized feature extraction techniques, the study demonstrates the efficacy of SER in real-world scenarios. The paper is structured to provide a review of related works in section 2, a methodology explanation in section 3, experimental results in section 4, and finally, in section 5, we conclude our work.

II. LITERATURE REVIEW

Numerous studies have examined speech emotion recognition. The success of numerous research projects will be reviewed here. Aharon et al. [1] used para-lingual information to train a deep neural network to detect speech emotions. The neural network learns emotional speech via recurrent and convolutional layers. The spoken signal spectrogram is used. Speech processing analyzes non-overlapping components. On IEMOCAP, this deep network and high-complexity convolutional LSTM algorithm achieved 68% recognition accuracy. Jonathan et al. [2] combined two machine-learning approaches for a better answer. They created unlabeled data with multitask machine learning and deep convolutional generative adversarial

networks. This expanded the speech emotions training corpus to 100 hours. Big data may improve speech emotion classifiers. The method-matching improvement was 43.88%. Chen et al. [3] suggested measuring deltas and delta-deltas for traits to maintain emotional information and reduce misclassification by reducing unnecessary components. Many SER frames are emotionless. Attention is amazing at considering complex feature representations. 3-D attention-dependent convolutional recurrent neural networks (ACRNNs) were trained to identify speech emotions using Mel spectrogram, deltas, and delta-deltas. Experimental results employing Emo-DB and IEMOCAP corpora show the proposed method has the highest unweighted average recall. Zhao et al. get log-Mel spectrograms and high-level features from raw audio[4]. Composite CNNs with 1D and 2D branches were built. Our combined deep CNN is built in two processes. Bayesian optimization selects training hyperparameters for the two architectural designs. After testing, the second thick layer was removed to integrate 1D and 2D CNN designs. Transfer learning accelerated merged CNN training. We started with 1D and 2D CNNs. 1D and 2D CNN features were blended to form CNN. Finally, the deep CNN with transferred functionality was optimized. Deep CNN fusion enhances emotion categorization on two benchmark datasets. Yenigalla et al. [5] assessed speech mood using phonemes and spectrograms. The phoneme sequence and spectrogram capture the expression's sentiment better than the text. Multiple deep neural network tests used phonemes and spectrograms. The essay uses a dataset to compare three network designs to cutting-edge approaches. These structures enhanced precision. The hybrid CNN model, which uses phoneme and spectrogram properties, was the most accurate IEMOCAP emotion model. The average class and total accuracy are higher than current methods. Using the IEMOCAP database, Sarma et al. [6] tested several DNN topologies for emotion recognition. Before training filters as network components, they compare time-domain and frequency-domain approaches of extracting functions using high-dimensional Mel-frequency cepstral coefficient (MFCC) input, similar to filter banks. The best filter-learning method is time-domain. The researchers next examined many speech data consolidation strategies. Their network investigations used temporal aggregation and single- or frame-by-frame speech labels. TDNN, LSTM, and time-restricted self-attention are their greatest designs. This design has 70.6% weighted precision, while the Fourier log-energy input with 257 dimensions technique has 61.8%. Latif et al. [7] improved Speech Emotion Recognition (SER) accuracy via cross-language and cross-corpus transition learning. DBNs outperform SVMs and sparse autoencoders for cross-corpus emotion identification. Five corpora in three languages show this dominance. The results also demonstrate that employing many languages during training and a tiny percentage of the target data enhances accuracy over the typical technique. This improvement applies to datasets with few training instances. Zhao et al. [8] introduced CNN+LSTM 1D and 2D networks. These networks use voice and log-mel spectrograms to learn local and global emotions. One LSTM layer and four LFLBs are in both networks. LFLB learns local and hierarchical correlations. One max-pooling and one convolutional layer. Our LSTM layer learns long-term dependencies from locally acquired functions. A sparse autoencoder and attention-focused technique were employed by Sun et al. [9]. An autoencoder learns from annotated and unannotated data, and the attention mechanism emphasizes emotional speech frames. Ignore emotionless speech frames. The approach is tested on three cross-language internet databases. Experimental results reveal that the suggested approach predicts speech emotion better than current algorithms. Using deep learning, Jiang et al. [10] extracted feature representations from distinct acoustic feature groups. This approach may include repetitious and inappropriate information, impairing emotion perception. Fusion networks learn discriminative acoustic feature representation and SVM as the final classifier after learning informative features. IEMOCAP studies indicated that the proposed design improved recognition efficiency by 64% over current methods. Pandey et al. [11] extensively studied deep learning algorithms for speech-based emotion extraction and categorization. Basic deep learning frameworks in papers are studied. Emo-DB and IEMOCAP used CNN and LSTM architectures to analyze Mel-spectrograms, magnitude spectrograms, and MFCCs for emotion capture. Analysis of testing results and rationale identified the best speech emotion recognition architecture and function coupling. A CNN, BLSTM, and SVM model can successfully categorize speech emotions using log-Mel spectrogram data, according to Zhen et al. [12]. The model outperforms another technique in prior studies using IEMOCAP. The model works well, but more data is needed to verify its applicability. However, [13] demonstrated the efficacy of different Speech Emotion Recognition (SER) models on six speech datasets. CNN+LSTM performs best in five of six datasets. Lili Guo et al. [14] categorized vocal emotions using KELM. Combining spectral characteristics trains the model. The model is tested using Emo-DB and IEMOCAP. The approach performs well on a specific dataset, suggesting it is not universal. Scientists found spectrum characteristics boost model categorization accuracy. Misbah et al. [15] extracted properties from raw speech log-mel spectrograms using a DCNN. Study datasets included IEMOCAP, Emo-DB, SAVEE, and RAVDESS. Four classifiers—SVM, random forest, k closest neighbors, and neural networks—classify speech emotions. These classifiers showed potential, but none worked on all four datasets. These classifiers may not generalize.

III. Methodology

The main purpose of SER is to increase human-machine interaction. To create a model for speech emotion recognition, many datasets are used.

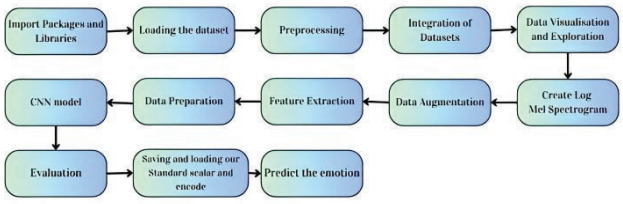

Fig. 1. Block Diagram

Working Model:

To begin with, it is crucial to remember that there are numerous platforms available for running and executing Python code. Several platforms are available, including

1. Google Colab
2. Kaggle

3. Pycharm

4. Jupyter Notebook

This study mostly focuses on Kaggle. Kaggle Notebooks is a cloud-based computing tool designed to support collaborative and repeatable analysis. The user has access to cutting-edge data analysis and machine learning programs that are pre-installed and compatible.

A. Loading the dataset

In this research on speech emotion recognition, the methodology is characterized by a systematic approach that starts with data imported from a RAVDESS dataset.

1) RAVDESS Dataset

In the 24.8 GB Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), there are a total of 7,356 items. A total of 24 experienced actors—12 men and 12 women—translate two statements into a neutral North American spoken language for the database [19]. Speech includes a range of affective states, including surprise, disgust, fear, happiness, sadness, and anger. Similarly, the song incorporates these sentiments. Information such as the medium, speech channel, emotion, intensity of emotion, statement, repetition, and actor details can be categorized using filename identifiers during the process of organizing the dataset into a Pandas DataFrame.

2) CREMA DataFrame

CREMA-D [18] has 7,442 original clips from 91 actors. These clips featured 48 male and 43 female actors aged 20–74 from African American, Asian, Caucasian, Hispanic, and Unspecified backgrounds. 12 sentences were read by actors. The statements used one of six emotions (Anger, Disgust, Fear, Happy, Neutral, and Sad) and four emotion levels (Low, Medium, High, and Unspecified).

3) TESS dataset

TESS stands for Toronto emotional speech set [16] in which two actresses (ages 26 and 64) spoke a set of 200 target words with the phrase "Say the word _." These words were recorded as they showed each of seven emotions: anger, disgust, fear, happiness, happy surprise, sadness, and neutral. There are a total of 2800 data points, which are audio files. The information is set up so that each of the two female actors and their feelings are in their folders. That's where you can find the audio file with all 200 target words. The audio file is saved in the WAV format.

4) SAVEE Dataset

The SAVEE (Surrey Audio-Visual Expressed Emotion) database [17] was made by four guys who speak English as their first language. Their names are DC, JE, JK, and KL, and they are between the ages of 27 and 31 and are postgraduate students and researchers at the University of Surrey. Emotions like fear, anger, disgust, happiness, sadness, and surprise have been put into their groups by psychologists. That's what his cross-cultural studies [6] show, and that's also what most studies of automatic mood detection [12] looked for. Our team added neutral so that you can hear 7 different feelings. Three of the sentences were general, two were about emotions, and the other ten were about different emotions but had the same number of sounds. The book had 15 TIMIT sentences. The 30 neutral sentences were made up of the 3 common sentences and the 2 x 6 = 12 sentences that talked about feelings. Text of the Article This means that each person says 120 things, like, "She had your dark suit in dirty laundry water for a whole year." Anger: Who permitted the account to keep using money for expenses? I'm sick. Could you please take this dirty tablecloth to the dry cleaners for me? Do not be afraid. If you need medical help, call an ambulance. Happiness: Those musicians sound great together. Sadness: Every governor would rather not have to spend less. What we didn't expect was for the carpet cleaners to shampoo our oriental rug. Not sure: the best way to learn is to do extra work.

B. Preprocessing

The imported audio files are then pre-processed from the Ravdess, Crema DataFrame, TESS dataset, and SAVEE Dataset. Integration ensures consistency across datasets, and log mel spectrograms are created to succinctly represent audio signals for subsequent analysis.

C. Integration of Dataset

TABLE I. INSTANCES AND EMOTIONS

Name: Emotions, dtype: int64

disgust	1923
fear	1923
sad	1923
happy	1923
angry	1923
neutral	1895
surprise	652

Table I indicates that, for instance, there are 1923 instances labelled as "Disgust," 1923 instances labelled as "Fear," and so on.

D. Create mel Spectrogram

```
<matplotlib.colorbar.Colorbar at 0x7d69a9104550>
```

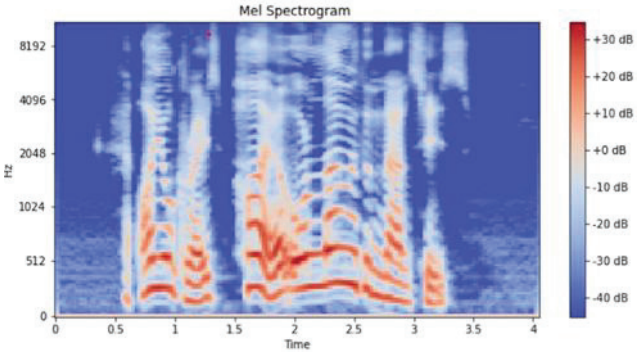

Fig. 2. Mel Spectrogram

The result of the Mel Spectrogram is shown in Fig. 2. Time is on the x-axis, frequency is on the y-axis (in the Mel scale), and color shows how strong the signal is at different time-frequency points. When working with speech and audio signals, this spectrogram is very helpful for seeing how the signal's frequency changes over time.

E. Data augmentation

Data augmentation techniques, including noise addition, stretching, shifting, and pitch alterations, are implemented to bolster the model's resilience.

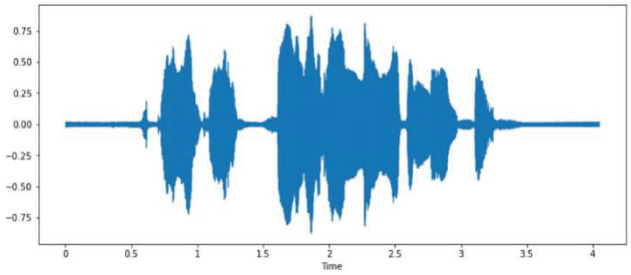

Fig. 3. Normal Audio

Fig 3 offers both a visual representation of the audio waveform and the ability to play the audio directly within the Jupyter Notebook or IPython environment.

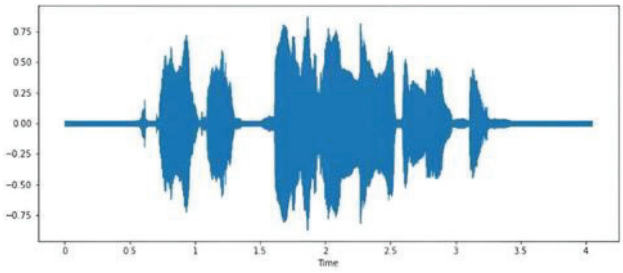

Fig. 4. Audio with Noise

Fig 4 provides both a visual representation of the noisy audio waveform and the ability to listen to the audio with added noise.

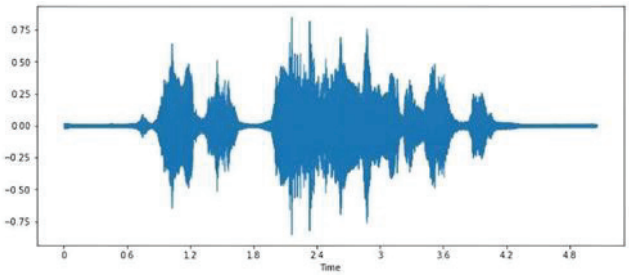

Fig. 5. Stretched Audio

Fig 5 provides both a visual representation of the stretched audio waveform and the ability to listen to the stretched audio.

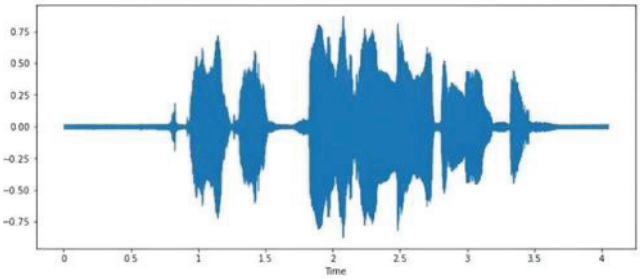

Fig. 6. Shifted Audio

Fig 6 provides both a visual representation of the shifted audio waveform and the ability to listen to the shifted audio. Shifting the audio in time introduces a time delay, which can be observed in the waveform plot and heard when playing the audio.

F. Feature Extraction

Feature extraction is performed on the pre-processed audio data using both standard and optimized methods to capture pertinent information for emotion recognition. Conventional and optimized approaches to feature extraction are explored, providing a baseline for comparison and improving computational efficiency.

Extracted features are saved to facilitate future experiments, and data preparation involves separating features and labels, applying one-hot encoding for multiclass classification, reshaping for compatibility with CNN models, and scaling using Sklearn's Standard Scaler. Early stopping is employed across all models during training.

	0	1	2	3	4	5	6
0	0.331543	0.471680	0.564941	0.452148	0.374512	0.296875	0.265137
1	0.238770	0.361816	0.478516	0.473633	0.485352	0.476074	0.472656
2	0.299805	0.419922	0.525879	0.459473	0.378418	0.326172	0.282227
3	0.252930	0.382812	0.497559	0.497070	0.487793	0.472656	0.482422
4	0.400879	0.591309	0.783203	0.777832	0.771973	0.777832	0.771973

Fig. 7. Features

2370	2371	2372	2373	2374	2375	Emotions
-1.234544	-0.693115	-0.038821	0.675410	1.405862	2.112551	surprise
1.766679	-1.586487	-0.501919	-3.159530	-5.015890	-0.942531	surprise
1.855116	2.404975	2.896071	3.282977	3.541091	3.666846	surprise
-3.266942	7.691891	7.443986	-2.031003	-2.095720	-1.418903	surprise
0.986784	-0.730886	-2.531058	-4.002848	-4.849192	-4.929412	neutral

Fig. 8. Features and Emotions

Fig 7 and 8 show the output with columns representing different features, and the 'Emotions' column containing the corresponding emotion labels. The head of the data frame is displayed, providing a glimpse of the data structure.

The research encompasses the implementation of a CNN model for speech emotion recognition, leveraging the pre-processed and scaled data. A Convolutional Neural Network (CNN) model is also developed, predictions are made on the test data, and random predictions are scrutinized for validation. Various plots are generated to visualize the performance of multi-models.

G. Predict the Emotions

TABLE II. PREDICTION TABLE

	Predicted Labels	**Actual Labels**
1.	angry	angry
2.	angry	angry
3.	disgust	disgust
4.	happy	happy
5.	fear	fear
6.	happy	happy
7.	fear	fear
8.	fear	fear
9.	surprise	surprise

Table II displays the predicted labels and actual labels for the initial 10 samples in the test set. It offers a concise summary of the model's performance on the test data. It is making predictions on the test data using a trained model. Itthen compares the predicted labels with the actual labels and creates a data frame to display the results.

H. Evaluation

Evaluation ensues, rigorously assessing the performance of different models, and results from the best-performing model are scrutinized based on predefined metrics. The best model is serialized to JSON and HDF5 formats, with the weights of the model also serialized for future use. Additionally, the Standard Scaler object used for scaling is saved, ensuring consistency in preprocessing. This methodology concludes with the development of a test script enabling the model to predict new records, showcasing its practical application in real-world scenarios.

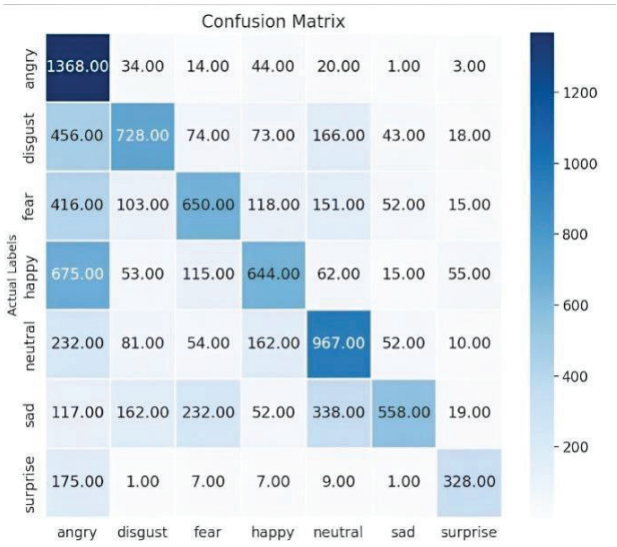

Fig. 9. Confusion Matrix

Figure 9 displays the result of the code, presenting a visual representation of the confusion matrix and a textual summary of the classification report. The confusion matrix heatmap provides a visual representation of the model's performance in accurately and inaccurately identifying cases for each emotion class. The categorization report offers further metrics to assess the model's performance on individual emotion classes.

TABLE III. PREDICTED LABELS

	Precision	Recall	F1-score	support
Angry	0.96	0.97	0.97	1484
Disgust	0.97	0.95	0.96	1558
Fear	0.96	0.97	0.96	1505
Happy	0.96	0.95	0.96	1619
Neutral	0.97	0.98	0.97	1558
Sad	0.96	0.97	0.96	1478
Surprise	0.98	0.97	0.97	528
Accuracy			0.96	9730
Macro avg	0.96	0.96	0.96	9730
Weighted avg	0.96	0.96	0.96	9730

Table III shows the classification report provides additional metrics for evaluating the model's performance on each emotion class.

IV. RESULT ANALYSIS

Speech emotion recognition analysis involves evaluating the performance of your machine learning model. This process includes dividing your data into training, validation, and testing sets, training the model, and assessing it using key metrics like accuracy, precision, recall, and F1-score. Visualizations such as confusion matrices provide insights into the model's performance. Error analysis helps identify patterns in misclassifications. Comparative analysis against benchmark models offers context. Optionally, examine bias and fairness. Qualitative analysis, listening to audio samples, confirms real-world applicability. Iterate to improve the model's accuracy and generalization, ensuring it aligns with human judgment in recognizing emotions in speech.

```
305/305 [==============================] - 7s 24ms/step - los
s: 0.1318 - accuracy: 0.9634
Accuracy of our model on test data :  96.34121060371399 %
```

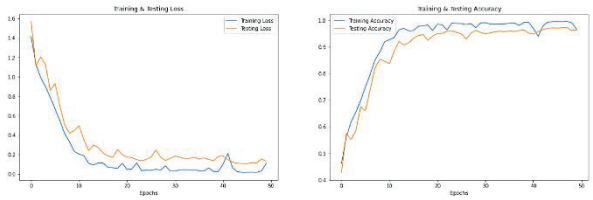

Fig. 10. Accuracy

Figure 10 displays the model's accuracy on the test data. Visual representations of the trends in training and testing loss and accuracy across epochs are the outputs. You can examine the model's performance in learning from the training data and in generalizing to the unseen test data from these charts.

V. CONCLUSION

In this research paper, our exploration has encompassed an array of models and methodologies, contributing to the collective understanding of SER systems and their potential applications.

The study combines and preprocesses audio data using a variety of datasets, such as RAVDESS, CREMA-D, TESS, and SAVEE, and uses log mel spectrograms to effectively extract features. Methods like CNN models are used, along with standard and optimized feature extraction techniques. The results show how good the suggested methods are at effectively distinguishing between speech and emotional signals, and they also show how important SER is in real-world applications. This research positions itself within the dynamic landscape of SER, contributing not only to the refinement of existing models but also to the identification of critical challenges that demand collective attention. To propel SER systems to new heights, concerted efforts are required to address these challenges. Future research endeavours could focus on the development of more diverse and inclusive datasets, encompassing a broader spectrum of languages, accents, and emotional expressions. Additionally, exploring novel feature extraction techniques, including the incorporation of contextual information and multimodal cues, may enhance the discriminative power of SER models. Interdisciplinary collaborations, involving linguists, psychologists, and computer scientists, can contribute valuable insights for a more comprehensive understanding of speech emotion dynamics.

REFERENCES

[1] M. RajaBabu, P. Abhinav, Nihaal Subhash—International Journal For Science Technology And Engineering-Vol. 11, Iss: 4, pp 1182-1185

[2] Sreeja Sasidharan Rajeswari1, G. Gopakumar1, Manjusha Nair1 Amrita Vishwa Vidyapeetham1-pp 169-178

[3] Anil Kumar Pagidirayi, Anuradha Bhuma -Revue Dintelligence Artificielle-Vol. 36, Iss: 2, pp 271-278.

[4] A. Kumar, V. Kumar and P. Rajakumar, "Speech Emotion Recognition Using Machine Learning," 2023 3rd International Conference on Innovative Practices in Technology and Management (ICIPTM), Uttar Pradesh, India, 2023, pp. 1-6.

[5] A. Kumar, V. Kumar and P. Rajakumar, "Speech Emotion Recognition Using Machine Learning," 2023 3rd International Conference on Innovative Practices in Technology and Management (ICIPTM), Uttar Pradesh, India, 2023, pp. 1-6.

[6] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. Ali Mahjoub, and C. Cleder, 'Automatic Speech Emotion Recognition Using Machine Learning', Social Media and Machine Learning. IntechOpen, Feb. 19, 2020. doi: 10.5772/intechopen.84856.

[7] K. V. Krishna, N. Sainath and A. M. Posonia, "Speech Emotion Recognition using Machine Learning," 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2022, pp. 1014-1018, doi: 10.1109/ICCMC53470.2022.9753976.

[8] Saw, A. K., Arya, C., Sahu, D., & Shrivas, S. (2022). Speech emotion recognition using machine learning. International Journal of Health Sciences, 6(S1), 14313–14321

[9] Amjad A, Khan L, Chang H. 2021. Effect on speech emotion classification of a feature selection approach using a convolutional neural network. PeerJ Computer Science 7:e766

[10] de Lope, J., Hernández, E., Vargas, V., Graña, M. (2021). Speech Emotion Recognition by Conventional Machine Learning and Deep Learning. In: Sanjurjo González, H., Pastor López, I., García Bringas, P., Quintián, H., Corchado, E. (eds) Hybrid Artificial Intelligent Systems. HAIS 2021. Lecture Notes in Computer Science (), vol 12886. Springer, Cham.

[11] S. Alisamir and F. Ringeval, "On the Evolution of Speech Representations for Affective Computing: A Brief History and Critical Overview," in IEEE Signal Processing Magazine, vol. 38, no. 6, pp. 12-21, Nov. 2021, doi: 10.1109/MSP.2021.3106890.

[12] H. J, R. R and R. S D, "A Study on Speech Emotion Prediction using Deep Learning Algorithm," 2021 Smart Technologies, Communication and Robotics (STCR), Sathyamangalam, India, 2021, pp. 1-5, doi: 10.1109/STCR51658.2021.9588861.

[13] H. Patni, A. Jagtap, V. Bhoyar and A. Gupta, "Speech Emotion Recognition using MFCC, GFCC, Chromagram and RMSE features," 2021 8th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 2021, pp. 892-897.

[14] A. Arun, I. Rallabhandi, S. Hebbar, A. Nair and R. Jayashree, "Emotion Recognition in Speech Using Machine Learning Techniques," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2021, pp. 01-07.

[15] C. Özkan and K. Oğuz, "Selecting Emotion Specific Speech Features to Distinguish One Emotion from Others," 2021 International Conference on Innovations in Intelligent Systems and Applications (INISTA), Kocaeli, Turkey, 2021, pp. 1-5.

[16] Pichora-Fuller, M. Kathleen; Dupuis, Kate, 2020, "Toronto emotional speech set (TESS)", https://doi.org/10.5683/SP2/E8H2MF, Borealis, V1

[17] S. Haq and P.J.B. Jackson, "Multimodal Emotion Recognition", In W. Wang (ed), Machine Audition: Principles, Algorithms and Systems, IGI Global Press, ISBN 978-1615209194, chapter 17, pp. 398-423, 2010.

[18] Keutmann, M. K., Moore, S. L., Savitt, A., & Gur, R. C. (2015). Generating an item pool for translational social cognition research: methodology and initial validation. Behavior research methods, 47(1), 228-234.

[19] Livingstone SR, Russo FA (2018) 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.https://doi.org/10.1371/journal.pone.0196391