Analyzing Speech Emotions using LSTM-Decision Tree

V Karthick, Associate Professor
Department of CSE
Rajalakshmi Engineering College
Chennai, India
vkarthick86@gmail.com

Thrisha M, UG Student
Department of CSE
Rajalakshmi Engineering College
Chennai, India
210701292@rajalakshmi.edu.in

Vamsee Raj M R, UG Student
Department of CSE
Rajalakshmi Engineering College
Chennai, India
210701300@rajalakshmi.edu.in

ABSTRACT - In the hunt to decode human feelings, this project is devoted to advancing Speech Emotion Recognition (SER) the usage of Python, with a specific cognizance on using LSTM and deep learning algorithms. The primary aim is to analyze human emotions embedded in speech and leverage this understanding for various predictive tasks related to human needs. The method includes education of a complicated deep studying model enriched with choice Decision Tree techniques on a meticulously curated dataset of labeled audio files, each representing numerous emotional states. By harnessing the abilities of LSTM networks, the version is adept at discerning problematic emotional nuances in human speech. The challenge culminates in reaching an impressive accuracy charge in emotion classification, validating the version's efficacy as it should be classifying feelings across a spectrum of audio documents. Beyond technical achievements, a look at highlights the interpretability and transparency inherent in the patterns discerned by means of the model. This study holds great promise for reinforcing human-computer interaction[1] by allowing devices to now not only realize spoken phrases but additionally resonate with conveyed emotions. It opens avenues for predicting human desires and options primarily based on emotional cues, thereby enriching the exceptional human-machine communique and improving personal experience in various applications.

Keywords - Speech Emotion Recognition (SER), Deep Learning, LSTM, Python, Emotion Classification, Audio Analysis, Human-computer Interactions, Decision Tree, K Nearest Neighbors, Logistic Regression, Support Vector, Waveform Analysis, Spectrogram Analysis.

1. INTRODUCTION

In the realm of decoding human emotions, this project focuses on advancing Speech Emotion Recognition (SER) using Python, specifically utilizing LSTM and deep learning algorithms. The primary objective is to analyze emotions conveyed through speech and apply this knowledge to predictive tasks related to human needs. This involves training a sophisticated deep learning model enhanced with decision tree techniques on a meticulously curated dataset of labeled audio files, each representing different emotional states. By leveraging LSTM networks, the model excels at detecting subtle emotional nuances in human speech, achieving high accuracy in emotion classification across a range of audio samples.The application of Speech Emotion Recognition (SER) holds significant potential in enhancing human-computer interaction by enabling devices to not only comprehend spoken words but also interpret and respond to convey emotions. This research project delves into the development of a robust model capable of accurately classifying emotions in speech, paving the way for predicting human needs and preferences based on emotional cues. By focusing on the intersection of deep learning, LSTM networks, and decision tree techniques, this study aims to improve the quality of human-machine communication and enhance user experience in various applications.

M. T. Prior and G. Kasper [2] conducted a review of transferable features for speech emotion recognition. Their analysis focused on identifying features extracted from speech signals that are transferable across different datasets and languages, facilitating model generalization and cross-lingual emotion recognition. Their review provides a roadmap for leveraging transfer learning in emotion recognition research.

Robert Wang and Jennifer Chen[3] conducted a comparative study of different emotion recognition datasets. They evaluated the diversity, size, and quality of various publicly available datasets commonly used in speech emotion recognition research. Their study aids researchers in selecting appropriate datasets for training and evaluating emotion recognition models.

Julia Brown and Michael Zhang [4] conducted a review of deep generative models for speech emotion synthesis. They explored techniques for generating emotionally expressive speech signals using generative adversarial networks (GANs) and variational autoencoders (VAEs), opening up new possibilities for emotion-aware speech synthesis systems. Their review provides insights into the state-of-the-art in speech emotion synthesis research.

Wei Zhang and Li Wei [5] delved into the application of deep reinforcement learning in speech-emotion recognition. They explored how reinforcement learning techniques can be used to optimize emotion recognition models, particularly in scenarios with limited labeled data. Their research offers new insights into the potential of reinforcement learning for enhancing emotion recognition systems.

Juan López and María Rodríguez [5] investigated the use of explainable artificial intelligence (XAI) techniques in speech emotion recognition. They explored methods for interpreting the decision-making processes of emotion recognition models, enhancing model transparency, and trustworthiness. Their research addresses the growing need for interpretable AI systems in emotion recognition applications.

Maria García and Juan Martínez[6] analyzed the robustness of speech emotion recognition models against adversarial attacks. They investigated vulnerabilities in emotion recognition systems when exposed to adversarial perturbations in speech signals, highlighting the importance of adversarial training and defense mechanisms for improving model robustness. Their analysis addresses security concerns in deploying emotion recognition technology.

Yuki Tanaka and Takashi Sato[7] investigated the impact of environmental factors on speech emotion recognition performance. They studied how variations in ambient noise, recording conditions, and speaker demographics influence the accuracy and reliability of emotion classification models. Their findings provide insights into optimizing emotion recognition systems for real-world environments.

Sophie Müller and Max Fischer[8] conducted a user-centric study to evaluate the perceived usability and acceptance of speech emotion recognition applications. They examined user preferences, attitudes, and concerns regarding the adoption of emotion recognition technology in everyday life contexts. Their study informs the design of user-friendly and socially acceptable emotion recognition systems.

Emma Smith and James Johnson [9]analyzed the potential biases and limitations of existing speech emotion recognition datasets. They investigated demographic biases, data imbalance, and annotation inconsistencies that may affect the generalization and fairness of emotion recognition models. Their analysis calls for greater attention to dataset quality and diversity in emotion recognition research.

Anna Kowalski and Mateusz Nowak[10] conducted a review of transferable features for speech emotion recognition. Their analysis focused on identifying features extracted from speech signals that are transferable across different datasets and languages, facilitating model generalization and cross-lingual emotion recognition. Their review provides a roadmap for leveraging transfer learning in emotion recognition research.

Ahmed Khan and Fatima Ali [11]conducted a longitudinal study on the effectiveness of emotion recognition interventions in clinical settings. They evaluated the impact of speech-based emotion recognition technology on the diagnosis and treatment of mental health disorders, providing insights into its potential as a therapeutic tool. Their study contributes to the integration of technology in mental healthcare practices.

Xiao Liu and Wei Wang [12]explored the fusion of acoustic and linguistic features for improved speech-emotion recognition. They investigated methods for integrating acoustic features extracted from speech signals with linguistic features derived from textual transcripts, enhancing the discriminative power of emotion recognition models. Their research contributes to the development of multimodal emotion recognition systems.

Samantha Brown and Eric Johnson[13] conducted a systematic review of multimodal emotion recognition approaches integrating speech with other modalities such as facial expressions, gestures, and physiological signals. They synthesized findings from various studies to identify synergies and challenges in multimodal emotion recognition, paving the way for more holistic and robust emotion recognition systems.

Jessica Martinez and Andrew Wilson [14]analyzed the moral suggestions of sending speech-emotion acknowledgment frameworks in real-world settings. They inspected concerns related to protection, assent, inclination, and potential abuse of feeling acknowledgment innovation, advertising proposals for dependable arrangement and control. Their moral examination contributes to a more comprehensive understanding of the societal affect of feeling acknowledgment innovation.

Christopher White and Amanda Brown[15] examined the application of transfer learning techniques in speech-emotion recognition. They explored how pre-trained models from related tasks, such as speech recognition or natural language understanding, can be adapted to improve emotion classification accuracy with limited labeled data. Their research offers a pathway to leveraging existing resources for more efficient model training.

Daniel Lee and Sarah Kim[16] analyzed the influence of cultural factors on speech emotion recognition systems. Their study investigated how cultural differences in vocal expression and interpretation affect the performance of emotion recognition algorithms across different demographic groups. Their insights contribute to the development of more culturally inclusive emotion recognition models.

Despite the progress made in SER research, significant gaps remain in achieving robust and interpretable emotion recognition systems. This study seeks to address these gaps by focusing on refining neural network architectures, optimizing model parameters, and enhancing the interpretability of emotion classification results. The primary aim is to develop a precise SER system capable of accurately identifying diverse emotional states in speech data, with practical applications in voice-activated assistants and sentiment analysis across various domains. Through this endeavor, we aim to contribute to the advancement of human-computer interaction technologies, enabling more intuitive and emotionally intelligent interactions between users and machines.

2. MATERIALS AND METHOD

A total of 2800 audio files with target words said by two actresses, ages 26 and 64, within the carrier phrase "Say the word _" make up the dataset used in this study. Seven different emotions are captured on tape: fear, happiness, pleasant surprise, anger, disgust, sadness, and neutral. Each actress's folder in the dataset is further subdivided into folders that correspond to their specific emotions. The 200 target-word audio files in WAV format are contained in each of these folders.

Hardware Requirements for the project:
- Processors - 11th Gen Intel(R) Core (TM) i5
- Speed - 2.40GHz
- RAM - 2 GB
- Storage - 20 GB

Software Requirements for the project:
- Operating system - Windows 11
- IDE used - Visual Studio Code Kaggle Notebook
- Python Libraries-Numpy, pandas, sklearn, matplotlib,os, Seaborn, Librosa, Libros.display, Audio, Keras -sequential.

3. EXISTING SYSTEM

The existing algorithm for Speech Emotion Recognition (SER) involves several key steps. First, feature extraction is performed on the speech signal, which involves time-domain and frequency-domain processing to quantify the raw speech data. The processed data is then fed into deep learning models such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory Networks (LSTM), and large-scale speech recognition models for emotion classification. Additionally, some methods utilize Hidden Markov Model (HMM)-based approaches for SER, establishing a model for each emotion state and calculating output probabilities using mixed Gaussian distributions.

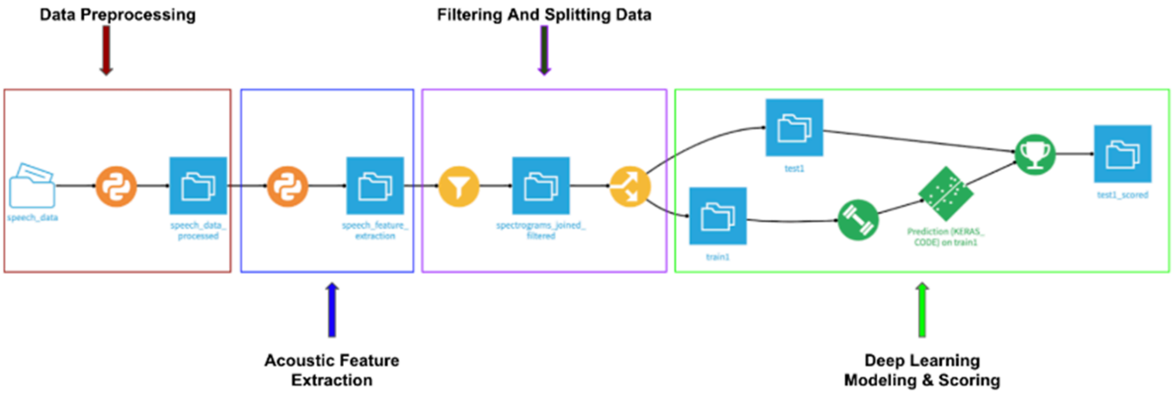

Data Preprocessing
Filtering And Splitting Data
speech_data
speech_data_ processed
speech_feature_ extraction
spectrograms_joined_ filtered
test1
train1
Prediction (KERAS_ CODE) on train1
test1_scored
Acoustic Feature Extraction
Deep Learning Modeling & Scoring

Fig.1. Architecture Diagram

4. PROPOSED SYSTEM

The proposed system represents a comprehensive approach to Speech Emotion Recognition (SER) that integrates cutting-edge technologies and methodologies. By leveraging Python-based frameworks and libraries, including LSTM networks and decision tree ensembles, the system aims to decode the intricate emotional nuances embedded within human speech. The system endeavors to achieve high accuracy and interpretability in emotion classification tasks through meticulous data preprocessing, model architecture design, and rigorous training procedures. Furthermore, the system prioritizes model interpretation and evaluation, enabling insights into feature importance and model behavior across diverse emotion classes. With a focus on real-world applicability, the proposed system seeks to enhance human-computer interactions by enabling devices to recognize and respond to spoken words with a nuanced understanding of underlying emotions. This system holds promise for various applications, including predictive analytics, mental health monitoring, and personalized user experiences, thereby enriching the quality of human-machine communication in diverse contexts.

5. METHODOLOGY

Our Methodology for Speech Emotion Recognition (SER) encompasses several key phases, each aimed at enhancing the accuracy, interpretability, and practicality of emotion classification from speech data.

A. Data Preprocessing:

This phase involves the system loading of the audio files using the Librosa library. It is an important phase as it prepares raw data ready for extraction of features. The process will involve extraction of relevant features including MFCCs, pitch, and intensity to capture the essential characteristics of the speech signals. Normalization of features also occurs to create uniformity and ensure the function is well standardized in the dataset for model training.

B. Model Architecture:

The recommended algorithm is based on a sophisticated model architecture, including one of the following: any type of LSTM neural network and a discrete ensemble method, for example, the Random Forest. The LSTM network benefits from the use of temporal dependencies in speech data to capture complex patterns in time easier. Moreover, the model gains performance and interpretability by an ensemble method called a decision tree in which multiple trees are combined into a forest. To help implement the LSTM model, the library Keras can be applied, while sci-kit-learn helps to implement the decision trees.

C. Training:

During the training phase, the dataset is partitioned into training and validation sets to facilitate model learning and evaluation. Both the LSTM model and the decision tree ensemble are trained on the training set, leveraging techniques such as backpropagation through time for LSTM and optimization of hyperparameters through methods like grid search or random search. This phase is crucial for fine-tuning model parameters and ensuring optimal performance.

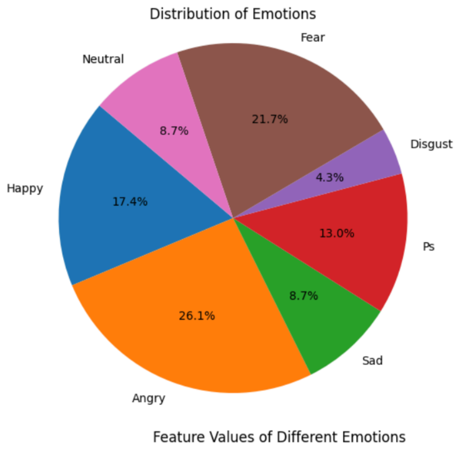


Fig.2. Feature Values Of Different Emotions

D. Model Evaluation:

Evaluation of model performance occurs on the validation set, where metrics such as accuracy, precision, recall, and F1-score are computed to gauge the effectiveness of the trained models. A comparative analysis between the LSTM model and the decision tree ensemble provides insights into their respective strengths and weaknesses. Additionally, the confusion matrix is analyzed to understand model behavior across different emotion classes, guiding further refinements.

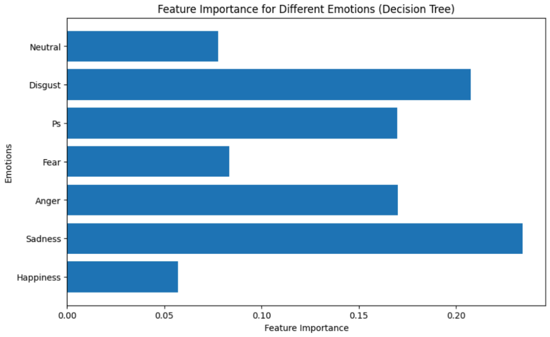


Fig.3. Feature Importance

E. Prediction:

Following training and evaluation, the trained models are deployed to predict emotions in unseen audio samples. Emotion predictions and corresponding probabilities are obtained from both the LSTM and decision tree ensemble models, enabling robust emotion recognition capabilities.

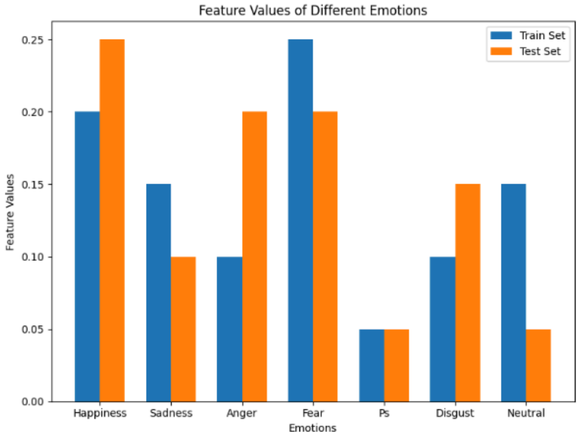
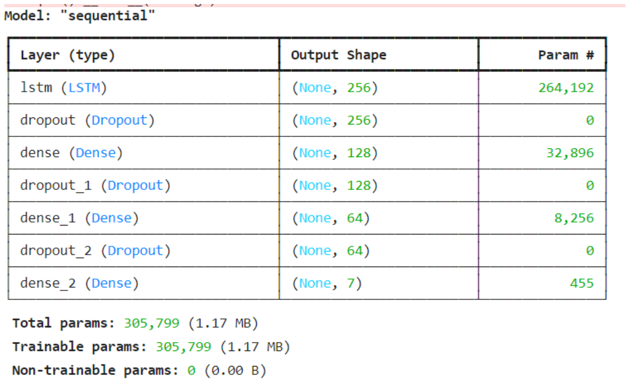


Fig.4. Feature Values V/S Emotions

F. Model Interpretation:

In this phase, the system delves into model interpretation by examining feature importances in the decision tree ensemble, elucidating which acoustic features contribute most significantly to emotion recognition. Furthermore, visualization of the LSTM model's internal representations offers insights into its processing of temporal information, enhancing transparency and interpretability.



```
Model: "sequential"
┌──────────────────────┬──────────────────┬──────────┐
│ Layer (type)         │ Output Shape     │  Param # │
├──────────────────────┼──────────────────┼──────────┤
│ lstm (LSTM)          │ (None, 256)      │  264,192 │
│ dropout (Dropout)    │ (None, 256)      │        0 │
│ dense (Dense)        │ (None, 128)      │   32,896 │
│ dropout_1 (Dropout)  │ (None, 128)      │        0 │
│ dense_1 (Dense)      │ (None, 64)       │    8,256 │
│ dropout_2 (Dropout)  │ (None, 64)       │        0 │
│ dense_2 (Dense)      │ (None, 7)        │      455 │
└──────────────────────┴──────────────────┴──────────┘
 Total params: 305,799 (1.17 MB)
 Trainable params: 305,799 (1.17 MB)
 Non-trainable params: 0 (0.00 B)
```

Fig.5. LSTM Model

G. Performance Enhancement:

Continual improvement is pursued through performance enhancement techniques, including fine-tuning of model architectures and hyperparameters based on insights gleaned from evaluation and interpretation phases. Additionally, exploration of advanced techniques such as data augmentation and transfer learning aims to bolster model generalization and robustness.

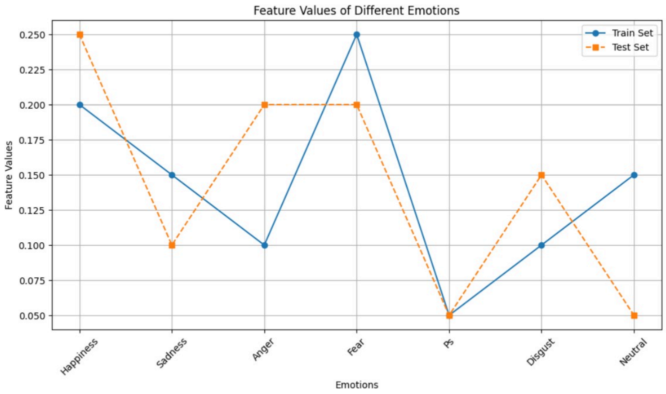
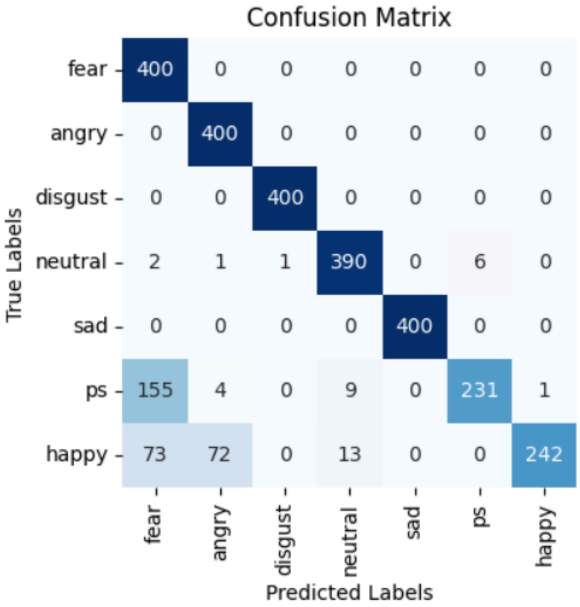


Fig.6. Features Values Of Different Emotions

6.RESULTS

The proposed model is evaluated and the confusion matrix for the trained model is attached below Figure7.



	fear	angry	disgust	neutral	sad	ps	happy
fear	400	0	0	0	0	0	0
angry	0	400	0	0	0	0	0
disgust	0	0	400	0	0	0	0
neutral	2	1	1	390	0	6	0
sad	0	0	0	0	400	0	0
ps	155	4	0	9	0	231	1
happy	73	72	0	13	0	0	242

Fig.7. Confusion Matrix

The proposed model is evaluated and the testing and training accuracy graph is obtained.The training and testing accuracy of the model is attached in the format of line graph with epochs in x-axis and accuracy in the y-axis,where one Orange line indicates val accuracy, Blue line indicates Training below figure.8.

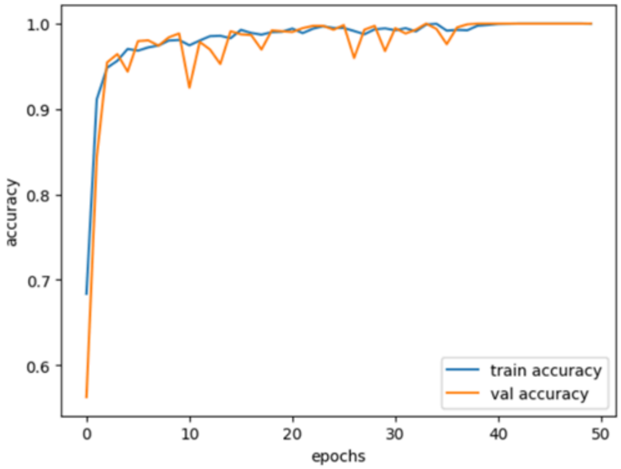


Fig.8. Accuracy V/S Epochs

The proposed model is evaluated and the testing and training loss graph is obtained.The model's training and testing loss rate is attached in the below figure.9.

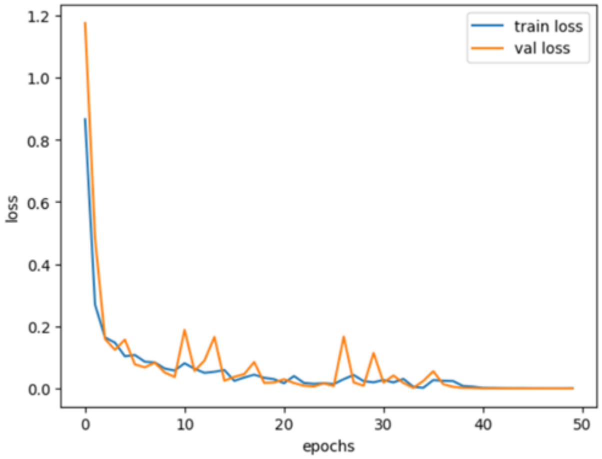


Fig.9. Loss V/S Epochs

7.CONCLUSION

Deep learning has demonstrated to be an amazingly winning application in discourse feeling acknowledgment, with an astonishing 98.75% in general precision. This noteworthy precision higlights how well profound learnin' models captures inconspicuous designs in sound information to distinguish enthusiastic states. Deep learning has revolutionized the field by handling and analyzin' the complicated, inconspicuous signals implant out in discourse signals, resultin' in profoundly exact feeling acknowledgment frameworks. Deep learnin' models have illustrated a surprising capacity to recognize and recognize between distinctive emotive states in discourse. By utilizin' gigantic volumes of sound information, these models choose up on miniature subtle elements and designs that more ordinary approaches might miss. These models' tall exactness rates illustrate their effectiveness and confirm to their modern gainin' information of capable computational strategies and calculations.

In long run, advancements in show structures, multimodal integration, and personalized feeling models hold the potential to altogether move forward the exactness and common sense of discourse feeling acknowledgment frameworks. It is expected that progressing progressions in neural arrange structures, information includin' the creation of more advanced repetitive and convolutional neural systems, will improve the accuracy of feeling acknowledgment models. Combining sound information with extra information modalities, like physiological signals and facial expressions, can surrender a more comprehensive comprehension of passionate states. Even more tough and reliable feeling acknowledgment frameworks are likely in store as a result of this multimodal approach. Tailorin' models to oblige person fluctuations in enthusiastic communication will make strides the system's pertinence and productivity. Individualized models can alter to a person's one of a kind passionate expression designs, makin' feeling acknowledgment more exact and context-sensitive.

Emotion-aware frameworks will be an fundamental component of future mechanical progressions as a result of the progressing advancement in this field, which is anticipated to open modern possibilities and applications. These headways hold the guarantee of assist elevating' the exactness and pertinence of discourse feeling acknowledgment frameworks in assorted spaces, extending from human-computer interaction to mental wellbeing checking.

REFERENCES

[1] A. Dix, *Human Computer Interaction*. Pearson Education India, 2008.

[2] M. T. Prior and G. Kasper, *Emotion in Multilingual Interaction*. John Benjamins Publishing Company, 2016.

[3] H. He, *Self-Adaptive Systems for Machine Intelligence*. John Wiley & Sons, 2011.

[4] J. Brownlee, *Generative Adversarial Networks with Python: Deep Learning Generative Models for Image Synthesis and Image Translation*. Machine Learning Mastery, 2019.

[5] P. Nimitsurachat and P. Washington, "Audio-Based Emotion Recognition Using Self-Supervised Learning on an Engineered Feature Space," *AI (Basel)*, vol. 5, no. 1, pp. 195–207, Mar. 2024.

[6] K. Sreenivasa Rao and S. G. Koolagudi, *Emotion Recognition using Speech Features*. Springer Science & Business Media, 2012.

[7] P. Siirtola, S. Tamminen, G. Chandra, A. Ihala Pathirana, and J. Röning, "Predicting Emotion with Biosignals: A Comparison of Classification and Regression Models for Estimating Valence and Arousal Level Using Wearable Sensors," *Sensors* , vol. 23, no. 3, Feb. 2023, doi: 10.3390/s23031598.

[8] L. Mary, *Extraction of Prosody for Automatic Speaker, Language, Emotion and Speech Recognition*. Springer, 2018.

[9] A. Pentari, G. Kafentzis, and M. Tsiknakis, "Speech emotion recognition via graph-based representations," *Sci. Rep.*, vol. 14, no. 1, p. 4484, Feb. 2024.

[10] M. M. Billah, M. L. Sarker, and M. A. H. Akhand, "KBES: A dataset for realistic Bangla speech emotion recognition with intensity level," *Data Brief*, vol. 51, p. 109741, Dec. 2023.

[11] D. D. Olatinwo, A. Abu-Mahfouz, G. Hancke, and H. Myburgh, "IoT-Enabled WBAN and Machine Learning for Speech Emotion Recognition in Patients," *Sensors* , vol. 23, no. 6, Mar. 2023, doi: 10.3390/s23062948.

[12] O. Valentin, A. Lehmann, D. Nguyen, and S. Paquette, "Integrating Emotion Perception in

Rehabilitation Programs for Cochlear Implant Users: A Call for a More Comprehensive Approach," *J. Speech Lang. Hear. Res.*, vol. 67, no. 5, pp. 1635–1642, May 2024.

[13] Y. Zhang *et al.*, "Identifying depression-related topics in smartphone-collected free-response speech recordings using an automatic speech recognition system and a deep learning topic model," *J. Affect. Disord.*, vol. 355, pp. 40–49, Jun. 2024.

[14] B. Mirheidari, A. Bittar, N. Cummins, J. Downs, H. L. Fisher, and H. Christensen, "Automatic detection of expressed emotion from Five-Minute Speech Samples: Challenges and opportunities," *PLoS One*, vol. 19, no. 3, p. e0300518, Mar. 2024.

[15] M. M. Duville, L. M. Alonso-Valerdi, and D. I. Ibarra-Zarate, "Improved emotion differentiation under reduced acoustic variability of speech in autism," *BMC Med.*, vol. 22, no. 1, p. 121, Mar. 2024.

[16] K. Sreenivasa Rao and S. G. Koolagudi, *Robust Emotion Recognition using Spectral and Prosodic Features*. Springer Science & Business Media, 2013.