International Journal of
Image Processing and
Pattern Recognition

ISSN: 2456-6985
Volume 10, Issue 2, 2024
July–December
DOI (Journal): 10.37628/IJIPPR

https://journalspub.com/journal/ijippr

Review IJIPPR

Face and Speech Emotion Recognition

Akshay Shetye[1], Shruti Chavan[2], Akshay Parab[2,*], Kaushik Patil[2], Sumitra Kulkarni[2]

Abstract

The study "faces and speech recognition" underscores the critical role of speech emotion recognition (SER) and its diverse applications across various fields, such as medicine, human-computer interaction, and customer service. SER has gained significant importance in cognitive psychology due to its potential to enhance user experience, improve patient care, and optimize the interaction between users and products. The process involves the identification and extraction of essential features from speech signals, which are key to accurately recognizing emotional states. The study further explores a range of classification algorithms that are employed to categorize these features, showing a transition from traditional AI techniques, like voice and energy-based analysis, to more advanced deep learning methods. These modern approaches utilize big data and neural network architectures to significantly improve the accuracy, reliability, and robustness of speech emotion recognition systems. Additionally, review articles in this field often draw upon data from SER research, offering valuable insights into the challenges and intricacies involved in the composition and implementation of such systems. The study traces the evolution of SER technology, emphasizing the benefits of deep learning, particularly its ability to learn directly from raw data, as well as the challenges posed by factors such as large log files and the need to accommodate various devices. Comprehensive reviews serve as crucial resources for researchers, practitioners, and policymakers, enabling a deeper understanding of the current state of SER technologies, including their strengths, limitations, and areas for future development. These insights are pivotal in driving the field forward, facilitating the development of innovative applications, and maximizing the potential of SER technologies in real-world scenarios.

Keywords: Feature extraction, machine learning, classification algorithm, natural language processing, speech emotion recognition

INTRODUCTION

Face and speech emotion recognition (SER) is a captivating area of research within the field of machine learning and artificial intelligence. It all comes down to how well machines can comprehend and interpret spoken language that expresses human emotions. The potential uses of this technology in a variety of fields, including entertainment, mental health diagnosis, customer service, and human-computer interaction, have drawn a lot of attention [1]. Our project's main goal is to use cutting-edge machine learning techniques to create a reliable and accurate SER system. Since emotions are a basic component of human communication, it may completely change how we interact with technology if we could automatically identify and interpret them from speech. The ability of apps to adjust to users' emotional states is one way that this system can improve the user experience. The gathering and categorization of a broad dataset of

***Author for Correspondence**
Akshay Parab
E-mail: parabakshay897@gmail.com

[1]Assistant Professor, Department of Computer Science and Engineering, Finolex Academy of Management and Technology, Ratnagiri, Maharashtra, India.
[2]Student, Department of Computer Science and Engineering (AI &ML), Finolex Academy of Management and Technology (FAMT), Ratnagiri, Maharashtra, India.

Received Date: June 28, 2024
Accepted Date: September 27, 2024
Published Date: October 08, 2024

Citation: Akshay Shetye, Shruti Chavan, Akshay Parab, Kaushik Patil, Sumitra Kulkarni. Face and Speech Emotion Recognition. International Journal of Image Processing and Pattern Recognition. 2024; 10(2): 35–43p.

speech samples expressing a variety of emotions, such as joy, sorrow, rage, and more, will be the focus of our study. After that, we'll use machine learning techniques like feature engineering and deep neural networks to train and improve our model. Our goal is to classify emotions with high accuracy while maintaining a robust system that can accommodate a variety of speaking tones and styles. We will explore the many facets of SER in this paper, covering topics such as the significance of emotion recognition, the methodology utilized, the dataset utilized, machine learning approaches, and assessment measures [2]. We will also talk about our SER system's useful applications and how they can affect human-computer interaction in the future. We will also examine the difficulties and potential paths for this developing profession, highlighting its advantages and disadvantages. The goal of our project is to make technology more sensitive to human emotions and enhance user experiences across a range of disciplines by advancing voice emotion recognition through machine learning [3].

PROPOSED METHODOLOGY

Our proposed methodology is intended to deliver an innovative solution for improving communication skills as shown in Figure 1.

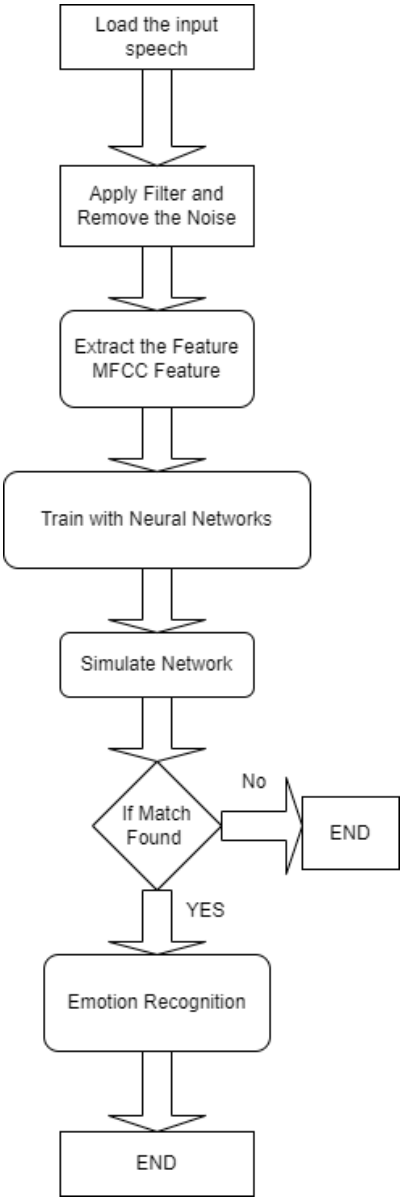

Figure 1. Workflow of model.

User Interface

- *User verification:* The system incorporates advanced facial recognition technology to authenticate users securely. This verification process ensures that only authorized individuals can access the system, thereby protecting sensitive data and personal information.
- *Facial emotion recognition:* The system uses sophisticated algorithms to detect and interpret a wide range of facial emotions, offering real-time feedback. This feature enhances user interaction by making the system more responsive to the user's emotional state.
- *Audio emotion analysis:* Users are prompted to record audio, which is then analyzed to identify underlying emotional tones. This feature allows for a more comprehensive understanding of the user's emotional state by considering both facial and vocal cues.
- *Get emotion analysis:* Users receive detailed reports on their emotional state based on their speech, including a sentiment rating and possible suggestions for emotional well-being. This analysis helps users gain insights into their emotional expressions [4].

Data Collection

- *Dataset gathering:* The methodology involves collecting a diverse and accurately labeled dataset of speech samples, each annotated with specific emotional labels. This dataset forms the foundation for training robust models capable of recognizing a wide range of emotions.
- *Dataset diversity:* Efforts are made to ensure that the dataset includes a variety of emotional expressions from different demographic groups, ensuring that the model is unbiased and can generalize well across different populations.

Datasets Utilized

- *Audio:* RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) provides high-quality audio and video recordings, which are essential for developing accurate emotion recognition models.
- *Facial data:* Datasets like AVEC2013 and AVEC2014 are used to train models on facial emotion recognition, offering a rich source of annotated facial expressions in various emotional contexts [5].

Data Cleaning

- *Audio processing:* The audio files undergo preprocessing steps, including down sampling and normalization, to ensure consistent quality. This process is crucial for minimizing noise and enhancing the clarity of the emotional cues in the audio.
- *Masking:* Masking techniques are employed to eliminate extraneous sounds and focus on the primary voice, improving the accuracy of emotion detection. This step is vital for ensuring that the model is not influenced by background noise.

Feature Extraction

- *Audio and facial features:* Key features are extracted from both audio files and facial data, including pitch, tone, intensity, and facial landmarks. These features are essential for accurately capturing the emotional nuances in the data [6].
- *Specific audio features:* The system extracts detailed audio features such as Mel-frequency cepstral coefficients (MFCCs), filter bank coefficients, and other relevant acoustic features, which are crucial for distinguishing between different emotions.

Label Classification

- *Emotion categorization:* The audio and facial data are categorized into predefined emotional groups, such as "happy," "sad," "angry," or "neutral". This categorization process involves sophisticated algorithms that analyze the extracted features to assign the most appropriate emotional label.
- *Feature analysis:* The classification process is enhanced by using advanced machine learning models, such as neural networks, which can capture complex patterns in the data [7].

Model Training

- *Dataset preparation:* The collected dataset is divided into training and testing subsets to evaluate the model's generalization ability. The training set is used to fine-tune the model's parameters, while the testing set provides an unbiased assessment of the model's performance.
- *Training:* The training phase involves iteratively adjusting the model's parameters to minimize the error in emotion recognition. Techniques such as cross-validation are used to ensure that the model does not overfit the training data [8].

Model Testing

- *Prediction:* The trained SER model is tested on unseen data to predict emotions. This step is crucial for assessing how well the model can generalize to new data.
- *Performance evaluation:* The model's predictions are compared with the actual labels to determine its accuracy and reliability. Metrics such as confusion matrices are used to visualize the model's performance across different emotion categories.

Model Evaluation

- *Comparison:* The model's predictions are rigorously compared to the actual labels in the test dataset to assess its accuracy and reliability. This comparison helps identify any biases or shortcomings in the model's predictions.
- *Performance metrics:* The model's performance is evaluated using a comprehensive set of metrics, including accuracy, F1 score, precision, and recall. These metrics provide a holistic view of the model's effectiveness and are crucial for identifying areas for improvement [9].

Real-World Testing

- *Application:* Once the model meets the desired performance criteria, it can be integrated into real-world applications such as sentiment analysis tools, voice assistants, and customer support systems. These applications benefit from enhanced emotional intelligence, allowing for more personalized and effective user interactions.
- *Utility:* The system's ability to recognize emotions from both speech and facial expressions provides a powerful tool for understanding and responding to human emotions in various contexts. This capability can be leveraged in fields, such as mental health, user experience design, and interactive entertainment.

RESULTS

Homepage

The homepage provides an intuitive user interface for accessing various functionalities of the emotion recognition system, offering a central point for navigation and user interaction, as shown in Figure 2.

Figure 2. Homepage.

Login Page

The login page ensures secure access to the system, requiring user authentication to protect sensitive data and personal information, as shown in Figure 3.

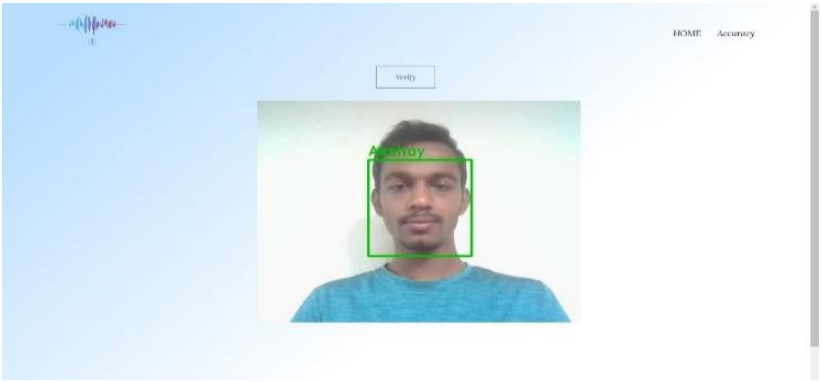

Figure 3. Login page.

Facial Emotion Recognition

Facial emotion recognition uses advanced computer vision techniques to analyze facial expressions and determine emotional states with high precision, as shown in Figure 4.

Figure 4. Facial emotion recognition.

Audio Emotion Analysis

Audio emotion analysis focuses on interpreting emotional cues from speech patterns, leveraging machine learning algorithms to classify emotions based on vocal features, as shown in Figure 5.

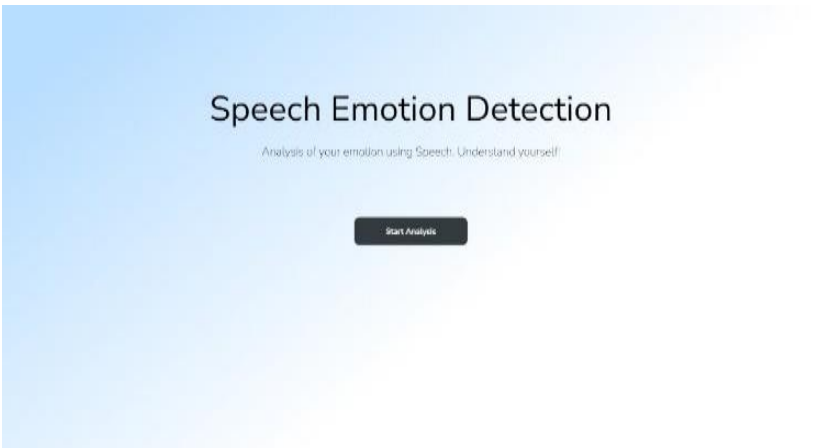

Figure 5. Audio emotion analysis.

Audio Record

Audio record feature allow users to capture and store audio samples for subsequent emotion analysis, enabling real-time or batch processing of voice data, as shown in Figure 6.

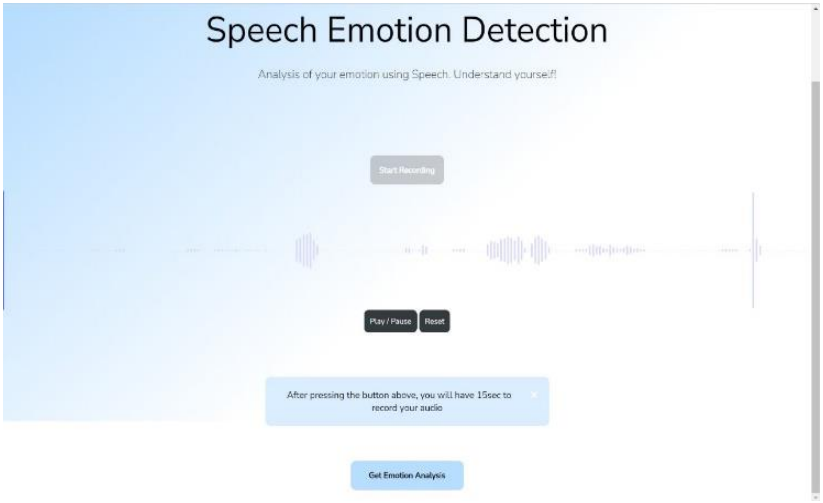

Figure 6. Audio record.

Emotion Analysis

Emotion analysis combines insights from both facial and audio data to provide a comprehensive understanding of user emotions, enhancing the system's ability to respond appropriately to emotional states, as shown in Figure 7.

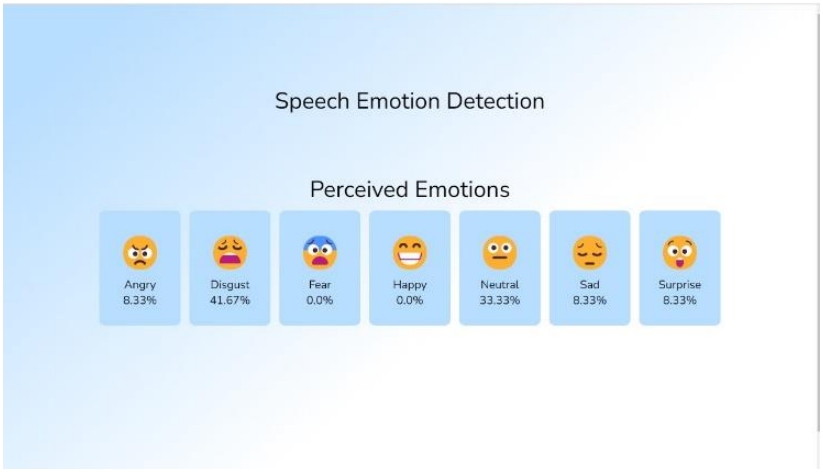

Figure 7. Emotion analysis.

BENEFITS

- *Improved human-computer interaction:* SER makes it possible for systems to recognize and react to human emotions, which enhances user-computer interaction [10].
- *Customer feedback analysis:* To learn more about the feelings and satisfaction levels of their customers, businesses can use SER to examine customer service calls and feedback.
- *Applications in healthcare:* By examining speech patterns for indications of emotional discomfort or problems, SER can help with mental health evaluations.
- *Education:* By evaluating students' emotional states and engagement levels, SER can be used in classroom settings to help customize instruction.
- *Call center monitoring:* Employers can enhance employee training and customer service by using SER to gauge the emotional content of calls made to contact centers [11, 12].

SCOPES

- *Subjectivity and variability:* Feelings are personal experiences, and various people may have different methods of expressing the same feeling. Furthermore, emotional expressiveness can be influenced by environmental and cultural circumstances, which can result in a variety of speech patterns.
- *Context dependency:* The conversational context frequently influences how people convey their emotions in speech. Various emotions can be expressed with the same words said in various situations. These contextual nuances may be difficult for current SER systems to comprehend effectively [13].
- *Multimodal signals:* Body language, gestures, and facial expressions are common examples of non-verbal signals used to convey emotions. When SER systems solely focus on speech analysis, they could overlook crucial clues offered by other modalities.
- *Limited datasets:* Large and diverse datasets are necessary for the development of accurate SER models. It can be difficult to compile these statistics with identified emotional content, though. Insufficient high-quality, well-annotated datasets could make it more difficult to train and assess SER systems.
- *Real-world noises and variability:* In real-world situations, a variety of environmental factors, including background noise, overlapping speech, and channel distortions, frequently alter speech signals. The performance of SER systems trained on sanitized and regulated datasets may be negatively impacted by these issues.
- *Imbalance in training data:* Natural language does not always convey emotions in the same manner. Imbalances resulting from the less frequent occurrence of certain emotions in the training set may impair the model's capacity to identify uncommon emotions [14].
- *Privacy concerns:* Since speech is a rich source of information about a person's emotional state, SER frequently entails analyzing sensitive personal data. A major challenge is ensuring the privacy and moral use of this data.
- *Interpersonal variability:* It is difficult to develop a universally applicable model for emotion identification since people exhibit their emotions in different ways depending on their unique personality features [15].

FUTURE IMPLICATIONS

Early detection of stress, anxiety, and other mental health indicators may be made easier with the integration of speech emotion recognition into wearable technology or smart settings, which could offer real-time emotional insights. Proactive interventions and individualized well-being support may result from this. By working together, SER and other AI technologies – like computer vision and natural language processing – may provide more complete multimodal emotion identification systems that advance our knowledge of complicated human emotions in a variety of settings. As these systems develop, resolving issues with algorithm bias, user permission, and possible abuse of emotional data will be necessary to ensure their broad adoption.

CONCLUSION

Face and speech emotion recognition (SER), a fascinating and rapidly expanding field, has the potential to alter how we interact with technology. As a result of our exploration of SER, the ability to identify and interpret human emotions from speech is no longer confined to the realm of science fiction but is instead rapidly becoming a reality. The value of SER goes far beyond its usefulness. It offers the chance to build machines that are more perceptive, sympathetic, and emotionally intelligent. Understanding the nuances of emotion in human communication opens the door to a wide range of practical applications, including customized customer service, mental health monitoring, educational assistance, and entertainment that can change to suit our moods. Even though SER has come a long way, it's important to acknowledge that there are still many challenges to be overcome. Challenges that must be overcome by ongoing research and innovation include the complexity of human emotion, cultural variations, and the need for enormous datasets. The ethical implications of privacy, consent,

and potential misuse of emotion data must also be carefully considered as SER technologies proliferate. A promising future lies ahead for face and speech emotion recognition as deep learning, machine learning, and artificial intelligence advances. Researchers are expanding the boundaries of what is conceivable, and with each new development, we inch closer to a future where technology can understand not only what we say but also our emotions. In essence, SER seeks to improve human-computer interaction rather than forge a stronger bond between people and technology. As SER technologies become more integrated into our daily lives, our interactions with machines may become more organic, intuitive, and emotionally engaging. Looking ahead, speech emotion recognition is clearly a game-changing force that will continue to transform how we interact with technology and communicate, ushering in a new era of human-computer symbiosis.

Acknowledgment

It gives the authors immense pleasure to present the paper on their project "Face and Speech Emotion Recognition" using machine learning. The authors would like to express their gratitude and indebtedness to those people who helped them in the completion of their project.

The authors owe their deep gratitude to their project guide, Prof. Akshay Shetye, who took keen interest in their project work and guided them all along, till the completion of their project work, by providing all necessary information. The authors are thankful and fortunate enough to get constant encouragement, support, and guidance from their teaching staff of the Department of Computer Science and Engineering (AIML), which helped them in successfully completing their project work.

REFERENCES

1. Kamińska D, Sapiński T, Anbarjafari G. Efficiency of chosen speech descriptors in relation to emotion recognition. EURASIP J Audio Speech Music Process. 2017;2017:1–9. doi:10.1186/s13636-017-0100-x.
2. Avots E, Sapiński T, Bachmann M, Kamińska D. Audiovisual emotion recognition in wild. Machine Vision and Applications. 2019;30(5):975–985. doi:10.1007/s00138-018-0960-9.
3. Baishya R. Unique solution of unpolarized evolution equations. Int J Res Appl Sci Eng Technol. 2020;8(4):499–509.
4. Poria S, Cambria E, Bajpai R, Hussain A. A review of affective computing: From unimodal analysis to multimodal fusion. Inform fusion. 2017;37:98–125. doi:10.1016/j.inffus.2017.02.003.
5. Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. Sci. 2017;356(6334):183–186. doi:10.1126/science.aal4230.
6. Cho J, Pappagari R, Kulkarni P, Villalba J, Carmiel Y, Dehak N. Deep neural networks for emotion recognition combining audio and transcripts. Interspeech. 2018;247–251.
7. Zheng L, Li Q, Ban H, Liu S. Speech emotion recognition based on convolution neural network combined with random forest. Chinese Control and Decision Conference (CCDC). Shenyang, China: 2018, Jun 9–11. 4143–4147. IEEE. doi:10.1109/CCDC.2018.8407844.
8. Weißkirchen N, Bock R, Wendemuth A. Recognition of emotional speech with convolutional neural networks by means of spectral estimates. Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW). San Antonio, TX, USA: 2017, Oct 23–26. 50–55. IEEE. doi:10.1109/ACIIW.2017.8272585.
9. Pandey SK, Shekhawat HS, Prasanna SM. Deep learning techniques for speech emotion recognition: A review. International Conference Radioelektronika (RADIOELEKTRONIKA). Pardubice, Czech Republic: 2019, Apr 16–18. 1–6. IEEE. doi:10.1109/RADIOELEK.2019.8733432.
10. Liu Y, Zhou M, Cao H, Liu H. Speech emotion recognition based on deep learning: A comprehensive survey. IEEE Trans Affect Comput. 2023;1–1.
11. Trinh Van L, Dao Thi Le T, Le Xuan T, Castelli E. Emotional speech recognition using deep neural networks. Sensors. 2022;22(4):1414. doi:10.3390/s22041414.

12. Lieskovská E, Jakubec M, Jarina R, Chmulík M. A review on speech emotion recognition using deep learning and attention mechanism. Electron. 2021;10(10):1163. doi:10.3390/electronics10101163.
13. Wani TM, Gunawan TS, Qadri SA, Kartiwi M, Ambikairajah E. A comprehensive review of speech emotion recognition systems. IEEE Access. 2021;9:47795–47814. doi:10.1109/ACCESS.2021.3068045.
14. Pan J, Fang W, Zhang Z, Chen B, Zhang Z, Wang S. Multimodal emotion recognition based on facial expressions, speech, and EEG. IEEE Open J Eng Med Biol. 2023;5:396–403. doi:10.1109/OJEMB.2023.3240280.
15. Han J, Zhang Z, Pantic M, Schuller B. Internet of emotional people: Towards continual affective computing cross cultures via audiovisual signals. Future Gener Comput Syst. 2021;114:294–306. doi:10.1016/j.future.2020.08.002.

