International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Face Emotion and Speech Recognition in Worker Stress Analysis

Prof. Sowndarya S[1], Ms. Ponlakshmi R[2], Ms. Tharani Priya M[3], Ms. Keerthilaya D[4], Ms. Priyanka Ss[5]

[1]assistant Professor, [2,3,4,5]student
[1,2,3,4,5]computer Science and Engineering
Knowledege Instiute of Technology

ABSTRACT

Human-computer interaction can benefit from better detection of emotion in the user. By allowing computers to detect and respond appropriately to users' emotions, user experience will be improved. A new project has been developed which uses both facial recognition technology and audio recordings to detect user emotion. The first part of the system uses facial emotion detection technology. A Haar cascade classifier is used to identify a face in an image or video feed before passing this data onto a CNN-based model (Mini-XCEPTION) that determines what kind of facial expression has been made (Anger, Disgust, Fear, Happiness, Sadness, Surprise, Neutral). The next part of the system uses audio detection technology to identify the speaker's emotional state by analyzing audio recordings or live audio streams recorded through a microphone. The system uses audio samples to compile relevant acoustic features (MFCC's, Pitch, Energy) that are analyzed by machine/deep learning algorithms to identify the speaker's emotion. Both facial and audio detection systems output data about your emotion, which is then processed together through a multimodal fusion mechanism to improve performance and accuracy when detecting user emotion. The entire system was created with Python and utilizes Stream lit to provide a user-friendly interface for displaying real-time visualizations, analyzing confidence levels while detecting emotion and tracking emotion overtime.

Keywords: Facial Emotion Recognition, Speech Emotion Recognition, Deep Learning

1. INTRODUCTION

In recent years, researchers have dedicated many resources to the study of emotion recognition within artificial intelligence as this could dramatically improve the interaction between people and machines via the understanding of emotions. David et al. (2018) suggest that many of the ways people will demonstrate their feelings will include behavioural and other patterns, such as facial expressions and speech, as two of the most popular and powerful indicators of a person's overall emotional state will include their facial expression and their speech signal. The traditional emotion recognition approaches using one form of detection may struggle to detect human emotions correctly due to environmental factors, differences in facial expressions/speech, and/or background noise, thus it is likely that emotion recognition would benefit

from a combination of multiple modalities to improve on their overall reliability/accuracy. Through the use of deep learning, machine learning and signal processing technique advancements, researchers are currently using automated methods to identify relevant feature data from a subject's facial expression and speech signal in order to classify the emotions of human beings. Therefore, in this case, the proposed system is a multimodal approach to emotion classification that combines facial emotion identification and speech emotion identification to help individuals evaluate human emotions more effectively.

A. Facial Emotion Recognition

Facial Emotion Recognition (FER) is important for different kinds of computer vision programs. The main focus for FER programs is to automatically identify what type of Emotion a person is feeling based solely on their facial expression from images or video files. The first step when processing an image of a person's face is to Locate the Face using an algorithm such as Haar Cascade Classifier to locate the face area in the input file, and then once Located, the Face is Extracted from the image, Pre-Processed by Resizing, Normalizing, and Converting into a Format that is suitable to be examined, and finally, the image file is processed by a CNN Based Deep Learning Program that has been Applied to Train to Recognize Different Expression Types. The training of the Deep Learning Network uses important facial features such as Eye Movement, Eyebrow Position, Mouth Shape, and Facial Muscle Patterns to create a Classification System that contains seven different types of Emotion or maybe just two: Angry, Disgusted, Scared, Happy, Sad, Surprised, and Neutral. FER will benefit from all computer programs that will enable humans to interact with their Computers In a much more effective manner by allowing Computers to Understand Emotions and then respond in a manner that is more appropriate to the Human Emotion.

B. Speech Emotion Recognition

spoken Emotion Recognition is a method that uses voice features to recognize and categorize human emotions from spoken data. Through changes in tone, pitch, intensity, and speaking patterns, human speech conveys important emotional information. During this procedure, the audio input is first recorded by the system using either an uploaded audio file or a microphone. To enhance the quality of the audio data, preprocessing techniques like noise reduction, silence removal, and normalizing are applied to the recorded voice signal. Important acoustic characteristics are retrieved from the speech signal after preprocessing, including pitch, energy, spectral patterns, and Mel-Frequency Cepstral Coefficients (MFCCs). These traits aid in recognizing emotional patterns and reflect the distinctive qualities of the voice.

C. Deep Learning

A branch of machine learning called "deep learning" focuses on employing multi-layered artificial neural networks to extract intricate patterns and representations from massive volumes of data. These neural networks allow computers to automatically learn features from raw data without the need for manual feature extraction because they are made to resemble the structure and operation of the human brain. Deep learning models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are frequently employed in emotion recognition systems to analyze audio and visual data. Because CNN models can automatically identify key facial traits like eye movement, mouth shape, and facial muscle patterns from photos, they are very good at recognizing facial emotions.

2. LITERATURE REVIEW

Suraj Tripathi [1] et.al has proposed in this system This study suggests a voice emotion recognition technique based on speech transcriptions (text) and speech attributes. While text aids in capturing semantic meaning, speech features like spectrograms and Mel-frequency Cepstral Coefficients (MFCC) aid in preserving emotion-related low-level qualities in speech, both of which support distinct facets of emotion detection. We conducted experiments using a number of Deep Neural Network (DNN) designs, which accept various text and speech feature combinations as inputs. On a benchmark dataset, the suggested network designs outperform state-of-the-art techniques in terms of accuracy. When it came to identifying emotions in IEMOCAP data, the combined MFCC-Text demonstrated the highest accuracy. Speech is required as input for most natural language processing technologies, including chatbots and voice-activated devices. It is standard practice to use Automatic Speech Recognition (ASR) systems to first convert this speech input to text, and then use the text output from the ASR to do categorization or other learning operations. Kim et al. [1] achieved state-of-the-art results on several benchmarks by training CNNs on top of pre-trained word vectors for sentence-level categorization. For text classification, Zhang et al. [2] employed character level CNNs and demonstrated results that were comparable to those of more conventional models like bag-of-words, n-grams and their TF-IDF versions, word-based Conv Nets, and recurrent neural networks (RNN).

Michael Neumann [2] et.al has proposed in this system One crucial and difficult task in the field of human-computer interaction is speech emotion recognition. A range of models and feature sets for system training were proposed in earlier work. In this study, we use a multi-view learning objective function and an attentive convolutional neural network to perform comprehensive experiments. We use various input signal lengths, acoustic feature kinds, and emotion speech (written or improvised) types to compare system performance. Regardless of the input feature selection, our experimental results on the Interactive Emotional Motion Capture (IEMOCAP) database show that the recognition performance is highly dependent on the type of speech data. Additionally, we obtained cutting-edge outcomes using IEMOCAP's improvised speech data. Recently, there has been a growing interest in speech emotion identification. This work is difficult since emotional expressions are complicated and influenced by a variety of characteristics, including age [1] and gender [2], and there is a dearth of large datasets. Deep learning (DL) has emerged as a cutting-edge technique for a variety of applications, including computer vision, speech recognition, and natural language processing (NLP). One unique type of neural network that has been effectively applied to both computer vision and speech [5, 6, 7] is the convolutional neural network (CNN), which was first introduced in [3, 4]. When it came to speech recognition, CNN outperformed other DL models in terms of noise resistance [8].

Rizwan Ullah [3] et.al has proposed in this system One of the most difficult tasks in human–computer interaction (HCI) systems is speech emotion recognition (SER). Effectively extracting emotional cues from a voice utterance is one of the main issues in speech emotion recognition. Even though recent studies have shown encouraging results, they often do not use sophisticated fusion algorithms to effectively express emotional elements in spoken utterances. We explain how to parallelize convolutional neural networks (CNNs) and a Transformer encoder for SER in order to fuse spatial and temporal feature representations of speech emotion. In order to simultaneously increase the filter depth and decrease the feature map with an expressive hierarchical feature representation at a lower computational cost, we stack

two parallel CNNs for spatial feature representation in parallel with a Transformer encoder for temporal feature representation. We identify eight distinct speech emotions using the RAVDESS dataset. To reduce model overfitting, we enhance and magnify the dataset's variances. To enhance the RAVDESS dataset, Additive White Gaussian Noise (AWGN) is employed. The SERVER model attains 82.31% accuracy for eight emotions on a hold-out dataset using the Transformer and CNNs' spatial and sequential feature representations. Furthermore, the SER system obtains 79.42% recognition accuracy for five emotions when tested using the IEMOCAP dataset.

Siddique Latif [4] et.al has proposed in this system Research on speech emotion recognition (SER) has historically depended on manually created acoustic features through feature engineering. However, creating bespoke features for intricate SER jobs takes a lot of manual labor, which hinders generalizability and slows down innovation. This has spurred the use of representation learning methods that do not require human feature engineering and can automatically learn an intermediate representation of the input signal. Faster innovation and enhanced SER performance are the results of representation learning. Deep learning (DL) advancements have further enhanced its efficacy by enabling deep representation learning, which automatically learns hierarchical representations from data. This work offers the first thorough analysis of the crucial subject of deep representation learning for SER. Humans naturally communicate by speaking to one another. It uses both explicit (linguistic) and implicit (paralinguistic) cues to communicate affective information regarding emotional expression. Language-dependent linguistic content and the difficulty of generalizing emotions across languages make linguistic communications somewhat unreliable tools for predicting and analyzing human affective behavior [1]. It might be challenging to predict a speaker's word choice and the related emotive responses because people frequently use diverse terms to convey emotion. Conversely, the paralinguistic content of speech offers a vast collection of auditory characteristics that can be utilized to encode the speaker's emotional state.

G. H. Mohmad Dar [5] et.al has proposed in this system In Human-Machine Interaction (HMI), emotion recognition from voice signals is essential, especially when developing applications like interactive systems and affective computing. With an emphasis on databases, feature extraction methods, and classification models, this review aims to offer a thorough analysis of contemporary approaches in speech emotion recognition (SER). In the past, techniques like Support Vector Machines (SVM), Random Forests (RF), and Gaussian Mixture Models (GMM) have used low-level descriptors (LLDs) like Mel-Frequency Cepstral Coefficients (MFCCs), linear predictive coding (LPC), and pitch-based features. However, the area has seen a radical transformation with the advent of deep learning techniques. Additionally, it examines how well various speech characteristics and classifiers handle issues including data imbalance, restricted data availability, and cross-lingual differences. In order to improve the performance of SE systems, the study emphasizes the necessity of further research into real-time processing, context-sensitive emotion recognition, and the integration of multi-modal input. This work attempts to offer a clearer route for improving feature extraction and classification methods in the field of emotion recognition by combining current developments and pointing out areas that require more investigation.

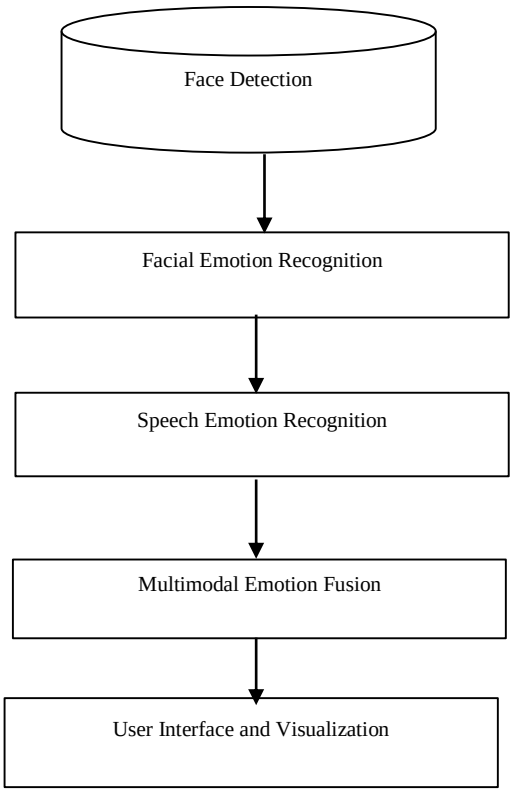

Fig 1 System Flow Diagram

3. PROPOSED SYSTEM

The proposed system comprises a multimodal emotion recognition solution that analyzes and identifies emotions as a result of analyzing human emotions from 2 different inputs: facial expression recognition & speech emotion recognition. Facial images will be captured by either a webcam or an uploaded image. The facial detection will use a Haar Cascade Classifier which will allow the system to detect the face region from the detected face region (image). The detected face will undergo preprocessing method processes as an emotional input & will employ a machine learning model (CNN) & a Convolutional Neural Network (CNN) to classify into predefined emotional expression categories like Angry, Disgusted, Fearful, Happy, Sad, Surprise, or Neutral. In speech recognition module, the microphone will either record the audio input or the recorded audio input will be used for analysis using speech recognition. After pre-processing using noise reduction and normalization methods, the acoustic features (MFCC, Pitch, Energy) will be extracted from each of these recorded audio input samples. These acoustic feature data will be processed using machine learning or deep learning algorithms to determine the emotional state represented in that audio input sample. At the end of both analysis' processes (facial & speech) using a fusion / combination solution, both outputs will be truly accurate & reliable representation of final emotions predicted.

A. Face Detection

Finding and identifying human faces in pictures or video frames taken with a camera is the responsibility of the Face Detection Module. The Haar Cascade classifier, a popular computer vision object detection algorithm, is employed in this module. The system uses trained characteristics to identify facial areas while

scanning the input image at various scales. The algorithm removes the facial region and gets it ready for additional processing after it detects a face. This stage is crucial because precise face detection guarantees that only the pertinent facial region is examined for emotion identification, enhancing the system's overall performance.

B. Facial Emotion Recognition

In order to determine the user's emotional expression, the Facial Emotion Recognition Module examines the identified face. Preprocessing involves scaling, normalizing, and transforming the retrieved facial image into a format that is appropriate for model input. The facial expression is then categorized into emotional groups such as Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral using a Convolutional Neural Network (CNN)-based deep learning model, such Mini-XCEPTION. The model determines the emotional state with a corresponding confidence score by analyzing face traits like eye movement, lip shape, and facial muscle patterns.

C. Speech Emotion Recognition

The emotional content of the user's speech is examined by the Speech Emotion Recognition Module. Preprocessing operations like noise reduction, silence removal, and normalizing are carried out by the system when audio input is recorded via a microphone or uploaded audio file. Following preprocessing, key acoustic characteristics are retrieved from the speech signal, including pitch, energy, and Mel-Frequency Cepstral Coefficients (MFCCs). The emotion expressed in the speech is then categorized by processing these features using machine learning or deep learning models. This module assists in identifying emotional states that might not be readily apparent from facial expressions alone.

D. Multimodal Emotion Fusion

The Multimodal Emotion Fusion Module combines the outputs from both facial emotion recognition and speech emotion recognition modules to produce a final emotion prediction. Since facial expressions and speech signals both carry emotional information, combining them helps improve the accuracy and reliability of the system. The fusion process is performed using a **confidence-based weighted mechanism**, where both modalities contribute to the final decision. This integrated approach reduces errors that may occur if only one modality is used for emotion detection.

E. User Interface and Visualization

Users can engage with the system and see the emotion detection findings on an interactive platform offered by the User Interface and Visualization Module. Stream lit is used to create the system, enabling users to record speech for analysis, submit photos, or record live video. The module shows confidence levels, detected emotions, and graphic charts that show the likelihood of each feeling. It also keeps track of emotional trends over time through analytical dashboards and emotion histories, which enhances the system's usability and informational value.

4. RESULTS ANALYSIS

This study implemented and evaluated a proposed Face Emotion Recognition System combined with Speech Emotion Recognition system. Both the facial emotion recognition and speech emotion recognition

modules successfully detected and classified the emotional state of the participant by analyzing their facial features, using Haar Cascade Classifier and CNN architecture, and extracting acoustic features (e.g. MFCC, pitch, energy) from the speech signal. The results of this study demonstrate that when facial and speech emotion recognition are integrated, there is an overall increase in detection accuracy compared to single-modality systems. Specifically, the multimodal fusion mechanism, which uses outputs from both systems to create a joint output, minimizes the number of incorrect classifications due to variations or fluctuations in facial expression or speech pattern. In addition, the results of the experiments showed that the system provided near real-time detection and visualization of participant emotions through a web-based interface developed using Stream lit, allowing users to easily understand and interpret the output of the study. Overall, the findings indicate that the system developed has the potential to accurately detect the emotional state of a participant and has numerous applications including use in mental health monitoring, providing feedback to humans who interact with computers, and developing emotion-aware intelligent systems.

1. Convolutional Neural Network (CNN)

CNN is used for **facial emotion recognition** to extract features from facial images.

Convolution Operation Formula

$$S(i,j) = (I * K)(i,j) = \sum_{m} \sum_{n} I(i+m, j+n)K(m,n)$$

Where:

- I= Input image
- K= Kernel / filter
- $S(i,j)$= Feature map output
- m, n= Kernel indices

2. ReLU Activation Function

Used in CNN layers to introduce **non-linearity**.

$$f(x) = \max(0, x)$$

Where:

- x= Input value
- $f(x)$= Activated output

This helps the network learn complex patterns in facial expressions.

3. SoftMax Function

SoftMax is used in the **output layer** of the CNN to convert scores into probabilities for each emotion class.

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}}$$

Where:

- $P(y_i)$= Probability of emotion class i
- z_i= Output score for class i
- k= Number of emotion classes

4. MFCC (Mel Frequency Cepstral Coefficient)

MFCC is used in **speech emotion recognition** to extract important audio features.

Mel Frequency Formula

$$Mel(f) = 2595 \times \log_{10}(1 + \frac{f}{700})$$

Where:

- f= Frequency in Hz
- $Mel(f)$= Frequency in Mel scale

MFCC represents the perceptual characteristics of speech signals.

5. Performance metrics

Accuracy

By determining the percentage of correctly classified samples (both positive and negative) relative to the total number of samples, accuracy gauges the model's overall correctness.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

Precision

Precision gauges how consistently accurate predictions are. It shows the proportion of anticipated positive samples that turn out to be positive.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall

The model's recall assesses how well it can recognize real positive cases. It displays the proportion of true positives that the model was able to capture.

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-Score

The harmonic mean of recall and precision is the F1-score. When working with skewed datasets, it offers a balanced metric that is particularly helpful.

$$\text{F1-score} = \frac{2.\ \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Table 1 Comparison Table

Model / Method	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Facial Emotion Recognition (CNN)	90.5	89.8	90.1	89.9
Speech Emotion Recognition (ML Model)	87.3	86.7	86.9	86.8
Multimodal Emotion Recognition (Face + Speech)	93.8	93.1	93.4	93.2

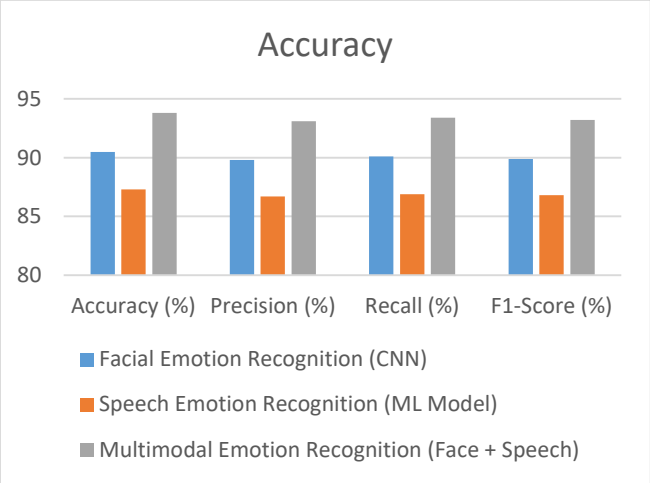

Fig 2 Comparison Graph

5. CONCLUSION

Overall, the proposed Face Emotion and Speech Emotion Recognition System allow users to detect emotion accurately and provide insight into human behaviour through facial expressions and speech signals. The integrated facial emotion recognition (using deep learning algorithms) and speech emotion recognition (acoustic feature analysis) provide a more accurate and reliable way to detect emotions than any one modality can do. By using a multimodal fusion method that integrates both modalities, this system provides a holistic view of emotion. The system was implemented using Python and Stream lit, providing a straightforward interface for users to detect real-time human emotions and visualize their results. Research indicates that the proposed system is capable of identifying the emotional state of an individual and therefore can be used to support applications such as mental health monitoring, human-computer interactions, virtual assistants, intelligent emotion-aware systems. Thus, the proposed system highlights the value of integrating different modalities to improve the overall effectiveness of emotion recognition systems.

6. FUTURE WORK

Larger datasets and more sophisticated deep learning architectures can be added to the suggested Face Emotion and Speech Emotion Recognition System in the future to increase the precision and resilience of emotion identification. The technology can be expanded to enable real-time emotion monitoring in online and mobile applications, increasing its usability. A more complete multimodal emotion detection framework can be created by integrating additional modalities as text sentiment analysis, body gesture recognition, and physiological data. Additionally, enhancing the speech processing module to manage various languages, accents, and noisy settings will make the system more useful in practical situations. In order to better understand individual emotional patterns and offer more accurate emotional insights for applications like mental health support, smart assistants, and human–computer interaction systems, the future system may also incorporate personalized emotion analysis and adaptive learning mechanisms.

REFERENCES

1. Satt, A., Rozenberg, S., & Hoory, R. (2017). Efficient emotion recognition from speech using deep learning on spectrograms. In Proc. Interspeech 2017, pp. 1089–1093.
2. Neumann, M., & Vu, N. T. (2017). Attentive Convolutional Neural Network Based Speech Emotion Recognition: A Study on the Impact of Input Features, Signal Length, and Acted Speech. In Interspeech 2017, pp. 1263–1267.
3. Zhao, S., Mao, X., & Chen, L. (2019). Speech Emotion Recognition Using Deep 1D & 2D CNN LSTM Networks. Biomedical Signal Processing and Control, 47, 312–323.
4. Latif, S., Qayyum, A., Usama, M., & Qadir, J. (2020). Survey of Deep Representation Learning for Speech Emotion Recognition. IEEE Access, 7, 123255–123287.
5. Akcay, M. B., & Oguz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication, 116, 56–76.DOI: 10.1016/j.specom.2019.12.001
6. S. Sadok, S. Leglaive, and R. Séguier, ‘‘A vector quantized masked autoencoder for speech emotion recognition,’’ 2023, arXiv:2304.11117
7. J. Singh, L. B. Saheer, and O. Faust, ‘‘Speech emotion recognition using attention model,’’ Int. J. Environ. Res. Public Health, vol. 20, no. 6, p. 5140, Mar. 2023
8. R. V. Darekar, M. Chavan, S. Sharanyaa, and N. M. Ranjan, ‘‘A hybrid meta-heuristic ensemble based classification technique speech emotion recognition,’’ Adv. Eng. Softw., vol. 180, Jun. 2023, Art. no. 103412
9. A. S. Alluhaidan, O. Saidani, R. Jahangir, M. A. Nauman, and O. S. Neffati, ‘‘Speech emotion recognition through hybrid features and convolutional neural network,’’ Appl. Sci., vol. 13, no. 8, p. 4750, Apr. 2023
10. M. J. Al Dujaili and A. Ebrahimi-Moghadam, ‘‘Automatic speech emotion recognition based on hybrid features with ANN, LDA and K_NN classifiers,’’ Multimedia Tools Appl., vol. 82, no. 27, pp. 1–19, 2023.