Emotions Recognition by Speech and Facial Expressions Analysis

Dr.G.Aparna[1], Bhukya Geethika[2], Bussa Sailahari[3], Vemula Ashwitha[4], CH. Uday Kiran[5]
[1]Associate Professor, *Hyderabad Institute of Technology and Management, Gowdavelli Village, Medchal, Hyderabad, India*
[2,3,4,5]UG Student, *Computer Science in Artificial Intelligence & Machine Learning, Hyderabad Institute of Technology and Management, Gowdavelli Village, Medchal, Hyderabad, India*

***Abstract*— Understanding human emotions is crucial in various real-world applications, such as mental health monitoring, human-computer interaction, and customer service automation. This project addresses the challenge of automatic emotion recognition by leveraging deep learning techniques on two key modalities: facial expressions and speech signals. Traditional methods often struggle with the variability of human emotions across different individuals and environments. To overcome this, we employ Convolutional Neural Networks (CNNs) for facial image and speech audio classification, allowing the system to learn discriminative features from both visual and auditory data.**

The facial CNN is trained on a dataset of facial images to identify emotions based on spatial features such as expressions, eye movements, and mouth shapes. Meanwhile, the speech CNN processes audio spectrograms to capture variations in tone, pitch, and rhythm that correlate with different emotional states. This approach enables a more reliable emotion recognition system by ensuring that emotions can still be detected even when one modality is unavailable or ambiguous. By addressing the growing demand for intelligent emotion-aware systems, this project contributes to fields such as healthcare (for stress and depression detection), customer service (for sentiment analysis in call centers), and smart assistants (for improving user experience through emotion-adaptive responses). The results lay the foundation for developing a more comprehensive multimodal emotion recognition system, bridging the gap between artificial intelligence and human-like emotional understanding.

Keywords: Emotion Recognition, Deep Learning, Convolutional Neural Networks (CNNs), Facial Expression Analysis, Speech Signal Processing, Multimodal Emotion Detection, Audio Spectrogram Classification, Mental Health Monitoring, Sentiment Analysis

I. INTRODUCTION

The ability to accurately perceive human emotions has long been a focus of research across various fields such as psychology, computer science, and artificial intelligence. In today's world, where human-computer interaction is rapidly evolving, creating systems that can understand and respond to human emotions has become increasingly significant. The project "Emotions Recognition by Speech and Facial Expressions Analysis" aims to build an intelligent system capable of recognizing human emotions by analyzing both vocal and facial cues. Emotions are inherently complex and can be expressed through a variety of non-verbal channels including voice tone, speech patterns, and facial expressions. Leveraging advancements in machine learning, computer vision, and audio signal processing, this project integrates multimodal data to improve the accuracy and reliability of emotion recognition. Speech emotion recognition is achieved by extracting important audio features such as MFCCs, chroma, and mel spectrograms, which are then fed into a convolutional neural network (CNN) to classify emotions like happiness, sadness, anger, and fear. Simultaneously, facial expression analysis is performed using deep learning techniques that detect and interpret facial muscle movements, identifying emotional states through visual data. By combining these two powerful sources of information, the system becomes more robust, compensating for limitations that might arise when only a single modality is used. The motivation behind this project stems from the growing need for empathetic technology in sectors like mental health care, customer service, education, and entertainment. Systems capable of understanding emotions can

enhance user experiences by offering personalized responses, detecting mental stress, or even assisting in therapy sessions. The project also addresses real-world challenges such as background noise in speech or occlusions in facial images by employing data preprocessing, augmentation, and careful model design. Through extensive experimentation and evaluation, the system aims to achieve high accuracy rates while being adaptable to diverse speakers and varying facial expressions. Overall, this project represents a significant step towards creating emotionally intelligent machines, bridging the gap between human emotion and technological response, and contributing to the broader goal of making human-computer interactions more natural, meaningful, and effective.

II.. LITERATURE REVIEW

Emotion recognition is a multidisciplinary field that combines aspects of psychology, linguistics, computer vision, and artificial intelligence. Over the years, researchers have made significant progress in developing systems that can automatically detect human emotions through various cues like speech, facial expressions, body posture, and physiological signals. Among these, speech and facial expressions are the most natural and prominent mediums for conveying human emotions.

Speech-Based Emotion Recognition

The task of recognizing emotions from speech has been studied extensively. Early approaches were primarily based on the extraction of prosodic features such as pitch, tone, rhythm, energy, and speech rate. These features were processed using classical machine learning techniques like Support Vector Machines (SVM), Gaussian Mixture Models (GMM), and Hidden Markov Models (HMM). These methods primarily relied on manually engineered features such as Mel-Frequency Cepstral Coefficients (MFCC) and spectral features, which are fundamental to representing the acoustic properties of speech signals.

However, despite these advancements, speech emotion recognition faced challenges related to speaker variability, language dependence, and environmental noise. To overcome these limitations, deep learning models began to emerge. Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks have shown remarkable success by learning meaningful features from raw audio data. These deep learning approaches are able to better generalize across speakers and languages, allowing for more accurate emotion detection, especially in noisy and variable conditions.

Facial Expression-Based Emotion Recognition

Facial expression recognition is another crucial domain within emotion recognition research. Initially, research focused on basic emotional categories such as happiness, sadness, anger, fear, surprise, and disgust. Early systems relied heavily on manual methods, such as landmark-based feature extraction techniques. The Facial Action Coding System (FACS) developed by Ekman and Friesen, which defines facial expressions in terms of muscle movements, played a significant role in understanding and categorizing facial emotions.

In recent years, deep learning techniques, particularly CNNs, have transformed facial emotion recognition. CNNs automatically learn hierarchical features from raw image data, enabling systems to detect subtle facial movements and changes with high accuracy. The use of modern neural networks like VGGNet, ResNet, and MobileNet has propelled facial expression recognition to new heights, offering improved performance over traditional handcrafted feature methods. However, challenges still persist, particularly regarding variations in lighting, occlusions (e.g., glasses or facial hair), and subtle emotions that are difficult to differentiate from one another.

Multimodal Emotion Recognition

Multimodal emotion recognition represents an advanced approach that combines information from both speech and facial expressions. While speech and facial expression recognition each have their own set of challenges, combining these modalities provides a more comprehensive understanding of emotional states. Multimodal systems are particularly useful because they can capture complementary features from both speech and facial expressions, improving the accuracy and robustness of the recognition process.

The fusion of speech and facial expression features can be performed at different stages. Feature-level

fusion involves combining the features from each modality before feeding them into a model. Alternatively, decision-level fusion involves integrating the outputs of separate unimodal models. A more sophisticated approach is hybrid fusion, which combines both feature and decision fusion for enhanced performance.

Deep learning models that utilize multimodal approaches include Multimodal Deep Belief Networks (DBNs), Multimodal LSTMs, and attention-based fusion networks. These architectures have shown promising results in recognizing emotions with higher accuracy, especially in real-world scenarios where both speech and facial expression data may be noisy or incomplete. Additionally, recent work in domain adaptation and transfer learning has contributed to generalizing multimodal emotion recognition models across different datasets and populations.

Current Trends and Gaps in the Field

Recent advancements in the field include the application of transformer models and self-attention mechanisms, which have been particularly effective in modeling long-range temporal dependencies in sequential data such as speech and facial expressions. Moreover, the integration of pretrained models and attention-based fusion strategies has led to significant improvements in both accuracy and efficiency. Despite these advancements, there remain several challenges that need to be addressed.

Real-time emotion recognition systems remain a difficult problem due to the computational complexity involved in processing both speech and facial expression data. Furthermore, recognizing emotions in natural, spontaneous conversations remains a significant hurdle, as individuals may express emotions subtly or in complex, dynamic ways. Another challenge is ensuring the ethical use of emotion recognition systems, particularly in applications such as surveillance, where concerns about privacy and misuse of personal data are critical.

In conclusion, the literature on emotion recognition reveals that multimodal systems, which combine speech and facial expression data, outperform unimodal approaches. The integration of deep learning techniques, particularly CNNs and LSTMs, has led to significant improvements in the accuracy and robustness of emotion recognition systems. Furthermore, advancements in multimodal fusion strategies and the adoption of state-of-the-art neural network architectures continue to push the boundaries of emotion recognition research. Building on these findings, this project aims to further explore and enhance the capabilities of multimodal emotion recognition by incorporating both speech and facial expression data, with the goal of achieving higher accuracy and better adaptability in real-world applications.

III. METHODOLOGY

Overview of the Proposed Approach--The methodology of this project revolves around building a dual-modality emotion recognition system that leverages deep learning techniques—particularly Convolutional Neural Networks (CNNs)—to recognize human emotions from both facial expressions and speech signals. The system is designed to handle data preprocessing, training, testing, and real-time prediction in a structured, step-by-step workflow.

Overview of Methodology

The development of an emotion recognition system that combines both facial expression analysis and speech signal processing is a complex yet rewarding endeavor. The proposed methodology follows a structured pipeline that starts from data collection and preprocessing, moves through training using Convolutional Neural Networks (CNNs), and ends with emotion prediction and performance evaluation. This hybrid approach ensures that the system can recognize emotions more reliably by leveraging both visual and auditory input channels. The following steps describe the methodology in comprehensive detail.

Dataset Upload and Initialization

The very first phase involves uploading datasets that contain labelled emotion data. For facial emotion recognition, this involves uploading a set of images where each image is tagged with a specific emotional label, such as “happiness,” “sadness,” “anger,” or “surprise.” These images should ideally feature a diverse range of subjects, expressions, lighting conditions, and angles to help the model learn more robust and generalizable features. Similarly, in the

case of speech emotion recognition, the dataset consists of recorded audio samples, also labeled with the emotion expressed in the voice. These datasets serve as the foundation for model training and validation.

Preprocessing of Facial and Speech Data

Once the datasets are uploaded, a comprehensive preprocessing step is essential. For facial image data, this begins with converting all images to a fixed resolution—typically 48x48 or 64x64 pixels. This uniform size helps maintain consistency across the dataset. Grayscale conversion is applied next, reducing the image channels and allowing the model to focus on structural and textural features such as mouth curvature, eyebrow orientation, and eye openness—features most relevant to emotional expression. Following this, pixel normalization scales the image data between 0 and 1, which accelerates learning by stabilizing gradient descent during training.

Data augmentation techniques are applied to artificially enlarge the dataset and prevent overfitting. Techniques like random rotations, horizontal flipping, zooming, and slight shifts enhance the model's ability to generalize to unseen expressions. After augmentation, the dataset is split into training and testing subsets using an 80:20 ratio. This means 80% of the data is used to train the CNN model, while the remaining 20% is used to test its performance after training is complete. For the speech data, preprocessing is slightly different but equally critical. Raw audio files are sampled and transformed into meaningful feature representations. Features such as Mel-Frequency Cepstral Coefficients (MFCCs), pitch, energy levels, zero-crossing rates, and spectral features are extracted because they effectively capture the nuances in human speech that reflect emotion. These extracted features are then converted into 2D spectrogram images, which are compatible with the CNN model and provide a visual form of the auditory signal that highlights variations in frequency and amplitude over time.

Training the Facial Emotion CNN Model

With the processed image data in hand, the next phase involves training the CNN model to recognize facial emotions. The CNN architecture typically includes several convolutional layers that automatically detect low-level and high-level features such as edges, shapes, and facial patterns. These are followed by pooling layers, usually max pooling, which reduce the dimensionality of the data while retaining key features. Activation functions like ReLU (Rectified Linear Unit) introduce non-linearity into the model, allowing it to learn complex mappings from input to output.

The final layers are fully connected dense layers that integrate all the learned features and feed into a softmax output layer. The softmax function calculates the probability of the input belonging to each emotion class, ultimately selecting the class with the highest probability as the model's prediction. The model is trained using a cross-entropy loss function, which measures the difference between the predicted and actual labels, and the Adam optimizer, which efficiently updates the model weights. This phase involves multiple training epochs, with validation accuracy and loss monitored after each epoch to avoid overfitting.

Training the Speech Emotion CNN Model

Parallel to the facial emotion model, a separate CNN is trained for speech-based emotion recognition. Instead of image pixels, the input consists of spectrograms or feature maps derived from audio signals. The CNN architecture for speech emotion recognition is similar to the one used for images but may include one-dimensional convolutional layers depending on whether the input is temporal (1D) or visual (2D). The purpose remains the same: to extract temporal patterns and frequency-based cues that indicate emotional states.

As with facial data, the model is trained on 80% of the speech dataset and validated against the remaining 20%. The training procedure includes monitoring accuracy, loss, and performance across different emotion classes. The speech model learns to identify subtle variations in tone, pitch, and energy that correlate with emotions like fear, excitement, or calmness.

Performance Evaluation and Accuracy Comparison

After training both models, a critical component of the methodology is evaluating and comparing their performances. Metrics such as accuracy, precision,

recall, F1-score, and confusion matrices are used to understand how well each model performs across different emotion categories. The results are visualized using an accuracy comparison graph, which provides a clear comparison of the facial CNN versus the speech CNN.

This step helps identify which modality performs better under specific conditions or for certain emotions. For instance, facial expression models may perform better at detecting joy or sadness, while speech-based models might better detect anger or stress. The evaluation also serves as a feedback loop for tuning hyperparameters or improving dataset quality in future iterations.

IV. MODEL AND ARCHITECTURE

The emotion detection system utilizes a multimodal Convolutional Neural Network (CNN) architecture, designed independently for speech-based and facial expression-based emotion recognition, followed by a fusion of their outputs to enhance prediction accuracy. The speech emotion recognition model accepts 2D feature representations of audio signals, primarily Mel-Frequency Cepstral Coefficients (MFCC), chroma features, and Mel spectrograms. These features are extracted using the Librosa library and stacked to form a feature map. The CNN model begins with one or more Conv2D layers that apply multiple filters (e.g., 32 or 64) with kernel sizes of 3×3 to learn local speech patterns. These are followed by MaxPooling2D layers to reduce dimensionality and retain the most significant features. Batch Normalization layers are used to normalize activations, enhancing training stability, while Dropout layers (typically set between 25%–50%) help mitigate overfitting. After convolutional processing, a Flatten layer converts the 2D feature map into a 1D vector, which is passed through one or more fully connected Dense layers. The final layer uses a softmax activation function to classify the input into predefined emotion categories.

In parallel, the facial expression recognition model takes preprocessed grayscale facial images as input, commonly resized to 48×48 pixels. This model also employs a CNN-based architecture with multiple stacked Conv2D layers to detect spatial features such as edges, facial landmarks, and expression patterns. As with the speech model, MaxPooling2D is used to downsample feature maps, followed by BatchNormalization and Dropout layers to optimize learning and avoid overfitting. The feature maps are then flattened and passed through Dense layers to output a probability distribution over emotion classes via a softmax layer.

After training both CNNs independently, the fusion of modalities is achieved by concatenating the output vectors (or intermediate feature representations) from the speech and facial models. This fused feature vector is further processed through additional Dense layers to perform final emotion classification. This multimodal approach leverages the complementary nature of speech and facial expression data, resulting in improved robustness and classification accuracy. The models are trained using the Adam optimizer and categorical cross-entropy as the loss function, with performance evaluated using standard metrics such as accuracy and loss on a separate test set. This architecture is capable of supporting real-time inference, enabling applications in healthcare, customer service, and virtual assistant systems.

INPUT AND OUTPUT DESIGN INPUT DESIGN:

The emotion recognition system uses two types of input data: speech audio and facial images. These inputs are collected, preprocessed, and converted into formats suitable for deep learning models. For speech, important features like MFCCs, chroma, and mel spectrograms are extracted. For facial images, grayscale conversion, resizing, and normalization are done to prepare the data.

The output of the system is an emotion label such as *Happy*, *Sad*, *Angry*, *Neutral*, etc. This label is generated after analyzing the input data using Convolutional Neural Networks (CNNs). When both speech and facial data are available, the system uses a multimodal approach by combining the results for higher accuracy.

This input-output design ensures the system can recognize human emotions effectively and be applied in areas like virtual assistants, mental health monitoring, and customer service.

INPUT DESIGN:

The input design of this project involves collecting and preparing two types of data: speech audio and facial images.

For speech input, we use audio files in .wav format. These audio files are preprocessed to remove noise and unnecessary parts. From these files, important features like MFCCs, chroma, and mel spectrograms are extracted. These features help the system understand the tone, rhythm, and pitch of the speech, which are important for recognizing emotions.

For facial expression input, we use images of people's faces. These images are converted to grayscale, resized to a fixed size (like 48×48 pixels), and the pixel values are normalized. This helps the model focus on important facial features like eyes, mouth, and eyebrows.

By properly preparing both speech and image data, the system can better learn and recognize different human emotions.

OUTPUT DESIGN:

The output of this project is the predicted emotion based on the input speech or facial expression data—or both when combined.

After processing the inputs through the trained models, the system gives an output in the form of an emotion label, such as:

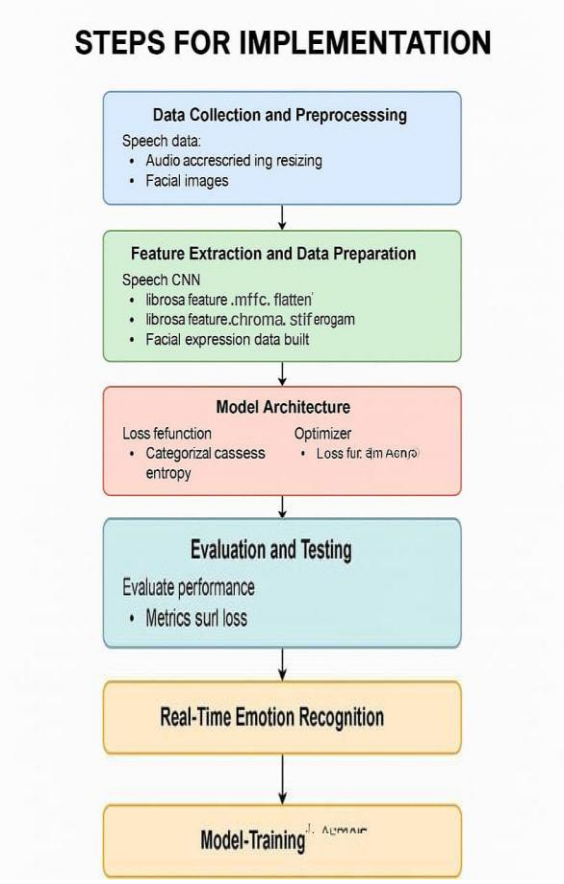
- Happy
- Sad
- Angry
- Fearful
- Neutral
- Disgust
- Surprise

This output can be displayed as text, shown on a webpage interface, or used in real-time applications like virtual assistants or healthcare monitoring systems. The system ensures that even if one input type (either speech or image) is unclear or missing, the other can still help in identifying the correct emotion.

The output is clear, easy to understand, and designed to help users or other systems respond appropriately based on the emotion detected.

V. IMPLEMENTATION

The implementation of this project focuses on Emotion Recognition through advanced Deep Learning techniques, utilizing Convolutional Neural Networks (CNNs) to analyze both Facial Expressions and Speech Signals for accurate emotional detection. The system is designed to process facial images by extracting key spatial features such as expressions, eye movements, and mouth shapes, allowing it to identify different emotional states. Concurrently, it processes audio data in the form of spectrograms, capturing tonal, pitch, and rhythmic variations that correspond to emotions. By integrating both visual and auditory data, the project aims to develop a robust Multimodal Emotion Detection system that can function effectively even when one modality is less reliable or unavailable.



This approach has wide-ranging applications, including Human-Computer Interaction (HCI), where it enhances user experience by enabling emotion-adaptive responses. It also plays a crucial role in Mental Health Monitoring by detecting signs of stress and depression through emotional cues. Additionally, the system is beneficial for Sentiment Analysis in

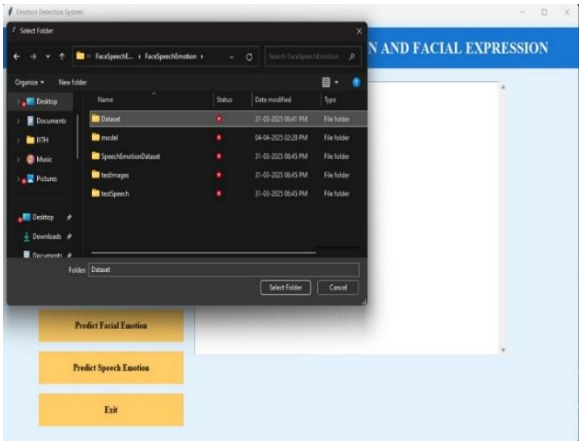
customer service settings, improving interactions by recognizing customer emotions. Through this comprehensive framework, the project contributes to creating more intelligent, emotion-aware systems.

.

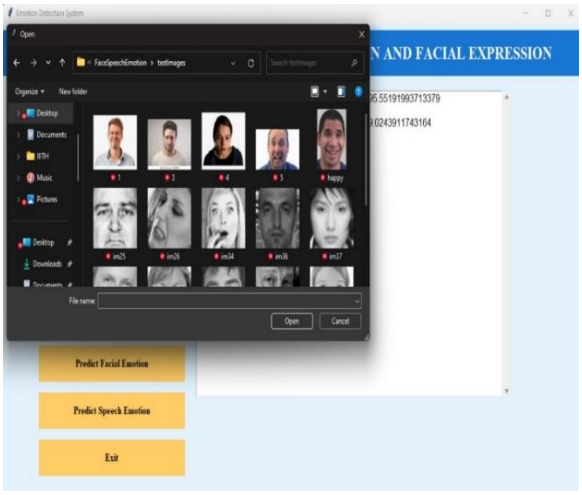
VI. TEST CASES AND FINAL RESULT

In this project we are detecting emotion using speech data and facial expression images and to implement this project we have trained CNN algorithm with RAVDESS Audio Dataset for speech emotion recognition and for face expression we have used Emotion Facial Expression images dataset. Below screen shots code with red colour comments showing extraction of MFCC features from audio dataset.

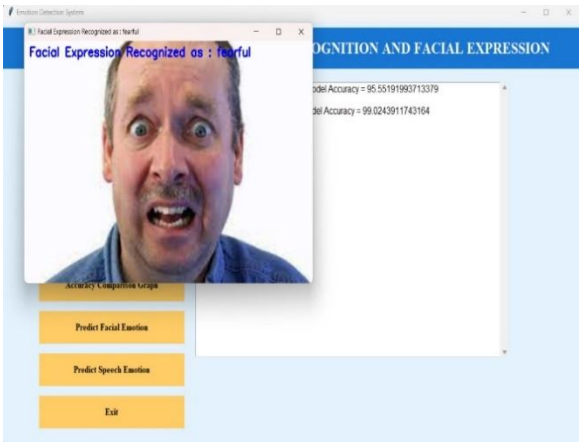
DATASET FOLDER TO BE LOADED:



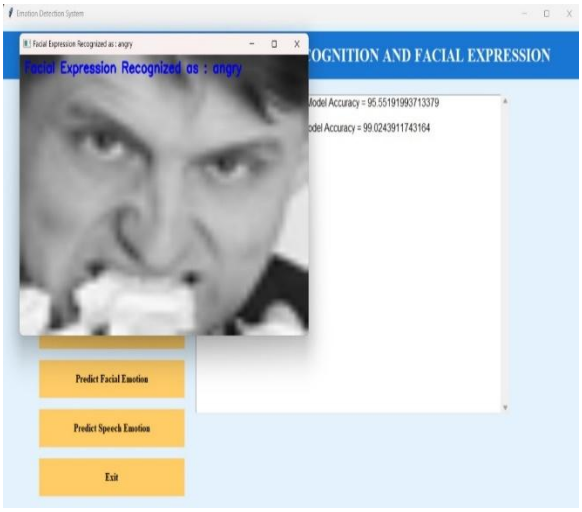
PICTURE FOR PREDICTING FACIAL IMAGES:



FACIAL EMOTION DETECTION: FEARFUL EXPRESSION IDENTIFIED:

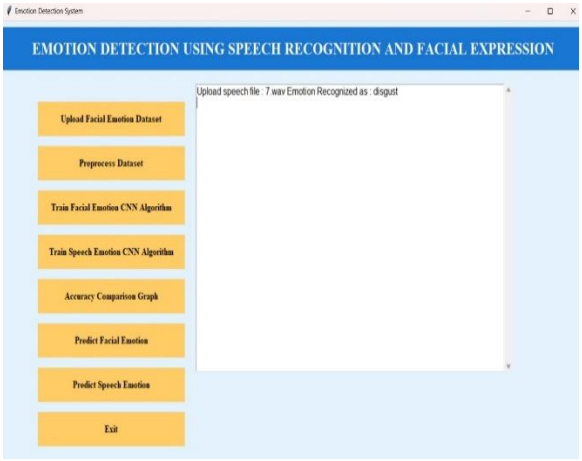


FACIAL EMOTION DETECTION SYSTEM IDENTIFYING ANGER:

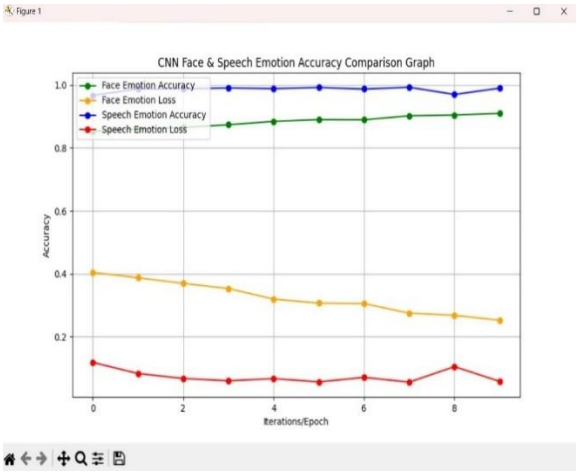


FACIAL-BASED EMOTION RECOGNITION SYSTEM

DETECTING DISGUST FROM AUDIO INPUT:



SPEECH-BASED EMOTION RECOGNITION SYSTEM DETECTING NEUTRAL FROM AUDIO INPUT:



ACCURACY:

The speech emotion recognition model consistently exhibits superior accuracy throughout the training process. It maintains an accuracy level close to 99%, reflecting high reliability in detecting emotions from audio features such as MFCCs, Chroma, and Mel spectrograms. In contrast, the facial emotion recognition model starts with an accuracy of approximately 88%, gradually increasing to around 93% by the final epoch. While still effective, it performs slightly below the speech-based model in terms of precision.

LOSS ANALYSIS:

Loss values further emphasize the difference in performance. The speech emotion model begins with a low loss value near 0.12, steadily reducing to below 0.08, indicating efficient convergence and strong generalization ability. Meanwhile, the facial emotion model shows a higher initial loss of around 0.42, which decreases gradually to 0.26. This trend indicates improvement in training but suggests the model requires more epochs or enhanced feature extraction for better convergence.

The comparative analysis demonstrates that the speech emotion model outperforms the facial emotion model in both accuracy and loss metrics. The audio-based system appears more robust in detecting emotional states, possibly due to the richer and more direct expression of emotion in voice signals compared to facial cues. This suggests that, under the current implementation, speech data provides more distinguishable features for accurate emotion recognition, and could be further leveraged in multimodal emotion detection systems.

VII. CONCLUSION

Emotion recognition gives an opportunity to significantly improve devices like cars, phones, TVs, office equipment and even household appliances and systems by implementing new features and interfaces, which would be much more intuitive and capable of auto adaptation to user needs. Large corporate companies use Affective computing, who want to know at all costs whether their products, services and marketing strategy addresses the needs and tastes of customers. Such technology can be researched and developed. In case of safety systems, for an example to know if the driver of any vehicle is active or not to know the dizziness of the driver. Which is already implemented by most innovative automotive manufactures. The great interest of this kind of research comes also from the Police and Security Forces, who see an opportunity for extraction a much larger amount of information recorded during interrogations and video surveillance. Future applications of such systems may be used in medicine, education or entertainment The polygraph is commonly known as a lie detector. A polygraph machine measures a physiological change in a person as a reaction to a mental thought. Polygraphs have been historically used to detect deceit, but those signals could be interpreted also for other mental states like anger, anxiety, depression etc. Developing emotion recognition systems, based on other signals than polygraph is using, gives an opportunity to transform polygraph into more robust and in the same time, less or even non invasive device. To build new technology for AI powered software to capture images in smart phone and smart cameras by facial expressions. Highly anticipated by the business market are emotion recognition solutions able to capture and analyze response and visual attention of consumers, compatible with ordinary web cams. It would certainly feed to the needs of Market Research, Brand Management, Creative Agencies and New Product Development. Helping make better decisions by incorporating customers emotions into their research. Game consoles allows as to play without any other controllers but our body, TVs are able to switch channels or change volume with single human gesture, and cameras will snap photo only when we smile. Extending those systems with the ability of

emotion recognition will let us to play a game on a console in a way like it was written just for us. TV could suggest which channel could be interesting for us for this day or even switch to other if we are bored. Detection of antisocial motives is currently given special emphasis for increased terrorist activity in our society. Emotional expression of the subject can be used to determine their possible anti-social motives. Emotion modeling has an interesting role in the next generation human machine interactive systems. It can be realized by modeling both input and output parameters of the interactive system.

REFERENCE

[1]. Bettadapura, V. (2012). Face expression recognition and analysis: the state of the art. arXiv preprint arXiv:1203.6722.

[2]. Shan, C., Gong, S., &McOwan, P. W. (2005, September). Robust facial expression recognition using local binary patterns. In Image Processing, 2005. ICIP 2005. IEEE International Conference on (Vol. 2, pp. II-370).IEEE.

[3]. Bhatt, M., Drashti, H., Rathod, M., Kirit, R., Agravat, M., &Shardul, J. (2014). A Studyof Local Binary Pattern Method for Facial Expression Detection. arXiv preprintar Xiv: 14 05.6130.

[4]. Chen, J., Chen, Z., Chi, Z., &Fu,H. (2014, August). Facial expression recognition based on facial components detection and hog features. In International Workshops on Electrical and Computer Engineering Subfields (pp. 884-888).

[5]. Ahmed, F., Bari, H., & Hossain, E. (2014). Personindependent facial expression recognition based on compound local binary pattern (CLBP). Int. Arab J. Inf. Technol., 11(2),195-203.

[6]. Happy, S. L., George, A., &Routray, A. (2012, December). A real time facial expression classification system using Local Binary Patterns. In Intelligent Human Computer Interaction (IHCI), 2012 4th International Conference on (pp. 1-5). IEEE.

[7]. Zhang, S., Zhao, X., & Lei, B. (2012). Facial expression recognition based on local binary patterns and local fisher discriminant analysis. WSEAS Trans. Signal Process, 8(1),21- 31.

[8]. Chibelushi, C. C., &Bourel, F. (2003). Facial expression recognition: A brief tutorial overview. CVonline: On-Line Compendium of Computer Vision,9.

[9]. Sokolova, M., Japkowicz, N., &Szpakowicz, S. (2006, December). Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In Australasian Joint Conference on Artificial Intelligence (pp. 1015-1021). Springer BerlinHeidelberg.

[10].Michel, P., & El Kaliouby, R. (2005).