iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IV **Month of publication:** April 2025

DOI: https://doi.org/10.22214/ijraset.2025.70092

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

Audio Aura-Speech Emotion Recognition System

Swayam Gawali[1], Ishaan G Koli[2], Suyog Gorule[3], Prathmesh Deokar[4], Rohini Deshpande[5],
Department of Computer Engineering Fr.Conceicao Rodrigues Institute of Technology Navi Mumbai, India

Abstract: ***Speech emotion recognition (SER) plays a crucial role in human-computer interaction, enabling systems to in- terpret and respond to user emotions effectively. In human- computer interaction, speech emotion recognition (SER) is es- sential because it allows systems to efficiently understand and react to user emotions. In this research, we introduce Audio Aura, a machine learning-based system for voice signal emotion classification. To improve classification accuracy and extractrich speech representations, the system uses a transformer-based model called Wav2Vec2. By leveraging Wav2Vec2's self- supervised learning capabilities, Audio Aura effectively captures temporal and contextual features in speech. The Toronto Emo-tionalSpeechSet(TESS)datasetisusedtotrainandassess the system, and it shows remarkable accuracy in recognizing emotions like neutrality, anger, sadness, and happiness. Com- pared to traditional machine learning approaches, transformer- based models demonstrate significant improvements in affective computing, making SER applications more robust in real-world scenarios.***
Index Terms: Speech Emotion Recognition, Deep Learning, Wav2Vec2,Transformer,SpeechProcessing,AffectiveComputing

I. INTRODUCTION

Speechisoneofthemostnaturalandexpressiveforms of human communication, conveying not only linguistic in- formation but also emotions, intentions, and sentiments. The ability to identify and interpret emotions from speech can greatly improve a variety of applications, such as virtual assistants, customer service automation, and human-computer interaction(HCI). Historically,emotionrecognitionhasrelied on textual sentiment analysis or facial expressions, but speech emotion recognition (SER) offers a more straightforward and unobtrusivealternative.MoreadvancedSERmodelshavebeen made possible in recent years by developments in artificial intelligence (AI) and deep learning. To determine a speaker's emotional state, these models examine changes in pitch, tone, intensity, and rhythm. However, issues like background noise, cultural differences, and inter-speaker variability still have an impact on how accurate and generalizable SER systems are.

This study introduces Audio Aura, a deep learning-based Speech Emotion Recognition system that can accurately clas-sifyemotionsfromspeechtoovercomethesedifficulties. Mel spectrograms and other acoustic features are among the meaningful representations that Audio Aura creates from raw speechdatausingafeatureextractionmethod.Theaudio data is then subjected to a Convolutional Neural Network (CNN)-based model, which effectively classifies emotions by capturing subtleties and complex patterns.The Toronto Emotional Speech Set (TESS) dataset, which includes high-quality emotional speech recordings from pro-fessional actors, is used to train and assess the model. The main objective of Audio Aura is to create an effective and scalable system that can identify a variety of emotions, in- cluding happiness, sadness, anger, surprise, and neutrality.

II. LITERATURE SURVEY

Speech Emotion Recognition (SER) systems are now cru- cial in improving human-computer interaction through the abilityofmachinestoperceiveandcomprehendemotions from speech. A real-time SER system based on Convolu- tional Neural Networks (CNNs) suggested by Attar et al. classified emotions through feature extraction in the form of Mel-Frequency Cepstral Coefficients (MFCCs). The system reached 80 percent accuracy using Ryerson Audio-Visual DatabaseofEmotional Speechand Song(RAVDESS)andhaspromisingpotentialforonlinelearning,robotics,andcustomer service[1]. Anotherproposalsuggestedby Deshmukhet al. employed a multilingual SER system with MFCC, pitch, and Short-Term Energy (STE) features and a Support Vector Machine (SVM) classifier. Their system reached 80 percent accuracy for American English, Hindi, and Marathi datasets, ensuring the robustness of the system and its application in multilingual settings [2]. The research by Gopikrishnan was based on the use of the Random Forest algorithm in emotion recognition, highlighting its efficiency and its potential for application in the healthcare, education, and customer service sectors. The paper also proposed the integration of biometric information and the enhancement of cross-cultural emotion recognition as future enhancements [3].Conversely, Bertero et al. created an interactive dialogue application emotion and sentiment recognition system in real- time. Using CNNs to directly extract emotion from the raw audio input, the system reached a precision of 65.7 percent in six classes of emotions and conducted sentiment analysis with an 82.5 F measure in transcribed speech using data from the out-of-domain[4].

©IJRASET: All Rights are Reserved

Wanietal.alsointroducedacomprehensive review of SER systems and compared conventional models such as GMMs with deep learning algorithms such as CNNs. The research indicated that CNNs considerably outperformed conventional techniques, although huge amounts of training data are needed. The problems of data set diversity and clas- sifier performance were introduced, with a recommendation for more multimodal and natural SER systems for real-world deployment [5]. In general, studies indicate robust feature extraction, sophisticated classification techniques, and cross- linguistic and real-time emotion detection to enable more user interaction and system flexibility.

III. PROPOSED SYSTEM

A. Problem Statement

Speech Emotion Recognition (SER) is a crucial aspect of human-computer interaction, enabling systems to detect and respond to users' emotions effectively. Traditional methodsof emotion recognition rely on text-based sentiment analysisor facial expressions, which may not always be accurate dueto contextual limitations and variations in human expressions. Additionally, existing speech-based recognition models either lack real-time processing capabilities or suffer from low ac- curacy due to inadequate feature extraction techniques.

To address these challenges, this research proposes a CNN- based Speech Emotion Recognition System, AudioAura, that effectivelyclassifiesemotionsfromspeechsignalsusingMel- Frequency Cepstral Coefficients (MFCCs) and deep learning techniques. This system aims to improve accuracy, robust- ness,andreal-timeapplicabilityinvariousdomainssuchas customer service, mental health monitoring, and human-robot interaction.

B. Scope

The proposed Speech Emotion Recognition (SER) system, AudioAura, aims to enhance human-computer interaction by enablingmachinestoperceiveandrespondtohumanemotions accurately. Unlike traditional text-based sentiment analysis or facial expression recognition, which may fail in audio-only scenarios, SER focuses on extracting meaningful emotional cues directly from speech signals. This system is designed to work efficiently in real-time, making it suitable for applica- tions where immediate emotional feedback is crucial, such as virtual assistants, customer support, and mental health moni- toring. By leveraging advanced feature extraction techniques, thesystemensuresthatvariationsinspeechpatterns,tone,and intensity are accurately captured, allowing for a more precise classification of emotions. The ability to recognize emotions across different languages and accents further broadens its usability in diverse environments.

Furthermore, the system's scalability allows it to be de- ployed in various sectors, including healthcare, education,and entertainment. In mental health applications, it can as-sist therapists in monitoring patients' emotional states during remote consultations. In call centers, it can analyze customer interactionsandprovidereal-timefeedbacktoenhanceservice quality. Additionally, it can be integrated into human-robot interaction, making robotic assistants more empathetic and responsivetohumanemotions.Withcontinuousadvancements in deep learning and audio processing, AudioAura has the potentialtoevolveintoamoresophisticatedsystem,capableofunderstandingnuancedemotionsandadaptingtouser-specific speechpatterns. Thisadaptabilitymakesitavaluabletool for improving communication between humans and artificial intelligence-driven systems.

IV. METHODOLOGY

The suggested Speech Emotion Recognition (SER) system has an organized workflow that allows for proper emotion classification from speech signals. The process involves four majorsteps:(1)DataPreparation,(2)Wav2Vec2Transformer FeatureExtraction,(3)RecognitionandClassification,and(4) Model Training and Testing. First, raw speech audio is preprocessed with methods like resampling, noise reduction, and data augmentation to improve its quality and maintain consistency across samples. The system then uses Wav2Vec2, a transformer-based self-supervised model, to obtain deep speech representations without the need for manual feature engineering. These features are then input to a Bidirectional Long Short-Term Memory (BiLSTM) neural network that learns sequence dependencies in speech data and classifies them into different emotion categories. Lastly, the system is optimized using an optimized learning strategy, and perfor- mance is measured using several metrics including accuracy, precision, recall, and F1-score. The following is the detailed methodology.

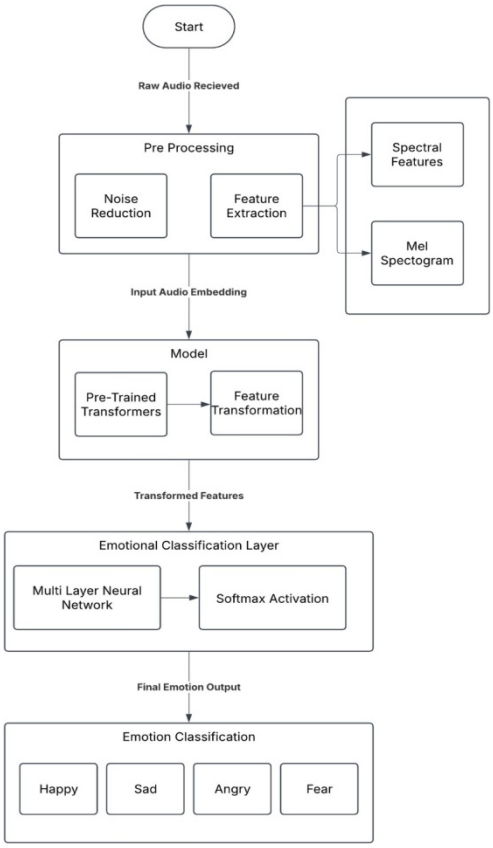


Fig.1.FlowchartoftheSystem

C. *Data Preparation*

1) *Data Selection:*An SER model's success relies uponthe diversity, quantity, and quality of training data. In this article, we experimented on widely used benchmark speech datasets such as RAVDESS (Ryerson Audio-Visual Database ofEmotionalSpeechandSong)andTESS(TorontoEmotional Speech Set) and finally we used TESS datasets. The datasets contain professionally recorded speech samples from multi- lingual speakers expressing various emotions such as Neutral, Happy, Sad, Angry, Fearful, Disgust, and Surprise. Pitch,tone,speechrate,andarticulationvariationsarecontrolled in each dataset to ensure rich coverage of natural emotional expressions in the real world. Having both male and female speakers renders the dataset suitable for deep learning-based emotion classification tasks. The use of multiple datasets ensures adequate generalization across emotional states and multiple speakers.
2) *Data Preprocessing:* To ensure consistent and high- quality input to the SER system, some preprocessing tech- niques were employed to pre-process raw speech signals. Resamplingwasperformedtonormalizeallaudiofilesto16 kHz,whichisthe Wav2Vec2modelrequirement.Noisereduc- tion technique such as spectral subtraction was employed to remove background noise and unwanted artifacts. Amplitude normalization was also performed to prevent volume bias by providing consistent volume levels for recordings. For model generalizationenhancement,dataaugmentationtechniquesuch as pitch shifting was employed, generating additional training samples to mimic real-world speech variations. These prepro- cessingtechniquesenhancethediversityandqualityoftraining data overall, resulting in a robust and accurate SER model.

D. *Feature Extractionusing Wav2Vec2Transformer*

1) *Self-Supervised Learning with Wav2Vec2:* Standard speechprocessingmethodsemployhand-craftedfeaturessuch as Mel-Frequency Cepstral Coefficients (MFCCs), pitch, and formants.However,deeplearning-basedmethods,particularly transformer-based models such as Wav2Vec2, are capable of extracting features automatically and hence handcrafting is avoided. The Wav2Vec2-large-960h model was employed in this research because it excels at learning significant speech patterns from raw audio waves. Contrary to typical practices, Wav2Vec2 employs self-supervised learning to learn features from speech without access to written transcripts.

The modelis trained on 960 hours of speech, and hence it is capable of learning speech sounds, rhythm, and timing. Employing this pre-trained model, the SER system learns useful speech infor- mation, leading to improved emotion classification accuracy compared to standard feature extraction methods.

2) *Feature Representation:*The raw waveform input is fed directlyintotheWav2Vec2model,anditgoesthroughaseries of transformer blocks and convolutional layers. The model outputscontextualizedfeatureembeddingsthatcapturenotable features such as intonation, rhythm, articulation, and stress patterns. In contrast to the previous methods involving fixed- window feature extraction, Wav2Vec2 is capable of learning torepresenttemporalrelationshipsinspeech.Theembeddings maintainthefinechangeintoneandpitch,whicharecriticalin the accurate identification of emotions. The extracted feature vectorsarefedintoadeeplearningclassifier,therebyensuring high-level speech representations are utilized effectively for emotion identification.

E. Classification and Recognition

Emotion Categories: The speech embeddings that are extracted are mapped to one of the recognized categories of emotions by a deep neural network classifier. The system recognizes these seven emotions:

- Neutral:Speakinanormalvoice,pace,andvolume, expressing no strong emotion.
- Happy: A more forceful tone, higher energy, and quicker speech, indicating happiness or excitement.
- Sad:Reducedspeechrate,reducedpitch,andmorepauses, conveying sadness or sorrow.
- Angry:Suddenchangesinvolume,quickspeech,and loud voice, employed while being angry.
- Fearful:Unsteadypitch,tremblingvoice,andirregular speech patterns, indicating anxiety.
- Disgust: Nasal resonance with irregular articulation, con- veying aversion or distaste.
- Surprise:Suddenshiftsintoneandrapidrisesinloud- ness, expressing surprise.

F. Classification Model Architecture

Wav2Vec2 embeddings from the audio signal were directly fed into a Softmax classifier to determine the emotional category. As opposed to models that are concerned with order in data, e.g., BiLSTM, the Softmax classifier can operate on featureembeddingsthatarecomplexinnatureandprovidethe probability score over various emotion classes. The best thing withthisapproachisthatWav2Vec2preservestimingrelations implicitly, and additional recurrent layers are not required to be used.
Thestepsfortheclassificationprocedureare:

- Feature Extraction: The Wav2Vec2 model, which has been pre-trained, transforms the raw sound wave into semantically meaningful vector embeddings.
- Flattening and Normalization: The embeddings collected are normalized and flattened to make features uniformly distributed.
- Fully Connected Layers: The embeddings are passed through fully connected layers with ReLU activation to assist in developing non-linearity and improved feature discrimination.
- Softmax Activation: The final layer employs a Softmax function.Thisonetransformsthelogitsintoaprobability distribution over the seven classes of emotions.
- Prediction:Theemotionlabelwiththehighestprobability score is taken as the final emotion label.

With Softmax classification, the system has a lightweight andefficientmodel,makingitsuitableforreal-timeSERappli- cations without compromising accuracy. The direct mapping from Wav2Vec2 embeddings to emotion classes enables the system to maintain high accuracy without having to cope with the complex calculations of recurrent networks.

G. Model Training and Evaluation

1) *Training Process:*The model was trained using super- vised learning, adjusting settings for optimal performance. Categorical Cross-Entropy was employed as the loss function asithelpslearnemotionclasseseffectively.AdamWoptimizer was selected since it can learn automatically to avoid overfit- ting. A cosine annealing learning rate scheduler was used to reduce the learning rate slowly, improving training. The batch sizewas16,findingtheperfectbalancebetweenbeingefficient and having the ability to generalize. The model was trainedfor 30 epochs on an NVIDIA GPU, leveraging hardware to speed up calculations.

©IJRASET: All Rights are Reserved

2) *EvaluationMetrics:*Toassessmodelperformance,mul- tiple evaluation metrics were used:
   - Accuracy:Measurestheoverallpercentageofcorrectly classified speech samples.
   - PrecisionandRecall:Evaluatetheproportionoftrue positives and false negatives.
   - F1-Score: Provides a balanced measure of precision and recall for each emotion class.
   - ConfusionMatrix:Visualizesmisclassificationsandhelps identify patterns in errors.
3) *Experimental Results:* The results show that the Wav2Vec2-Softmax model can differentiate between neutral, happy, sad, angry, fearful, disgusted, and surprised speech patterns. In addition, the model was effective even with different speaker accents, background noise, and recording environments. The combination of deep speech embeddings andaSoftmaxclassifierallowedforrapidemotionrecognition, which is useful in human-computer interaction, call center evaluation, and mental health monitoring.

V. SYSTEM IMPLEMENTATION

1) *System Workflow and Functionality:*The process of the Speech Emotion Recognition (SER) system is well-defined to parserawaudioandclassifyitintovariousemotions.Itbegins when users provide a speech sample through speaking into a microphone or loading an audio file. The captured or loaded audio is processed through some preparation stages, such as converting the sample rate to a common 16kHz frequency, noise reduction, and silence elimination to make the audio clear and uniform. Following the preparation of the audio, itis processed using the Wav2Vec2 transformer model, which isolates salient speech features that identify emotions.Once features are extracted, these embeddings are passed through a Softmax classifier, which outputs the probability scores for every pre-specified emotional category. The classi- fier projects the speech features extracted to any of the seven emotions: neutral, happy, sad, angry, fearful, disgusted, or surprised. The output is presented on an easy-to-use interface that shows both the extracted emotion and its confidence score. The system also permits data to be stored and analyzed further, allowing for insights into emotional trends across various users and interactions. The compact architecture ofthis system provides real-time processing with high accuracy and reliability.
2) *ImplementationTechnologies:*TheSERsystememploys different technologies in different domains, such as machine learning tools, backend development tools, and frontend de- sign. Python is the primary programming language employed tobuildmachinelearningmodels,usingtheWav2Vec2feature extractionmodelfromtheHuggingFaceTransformerslibrary. Deep learning frameworks such as PyTorch and TensorFlow are employed to train and fine-tune the model to enable it to learnappropriately.Audioprocessinglibrariessuchas Librosa, Soundfile,andPydubarealsoemployedtopreprocessthe speech signal prior to its use in classification.

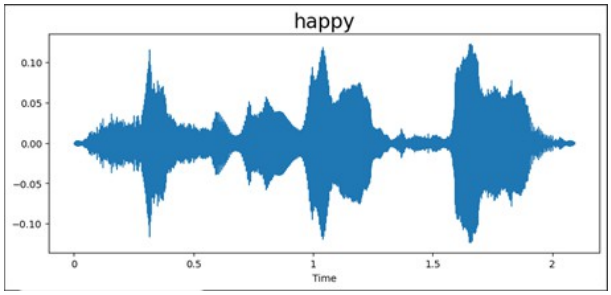


Fig.2.AudioWaveplot

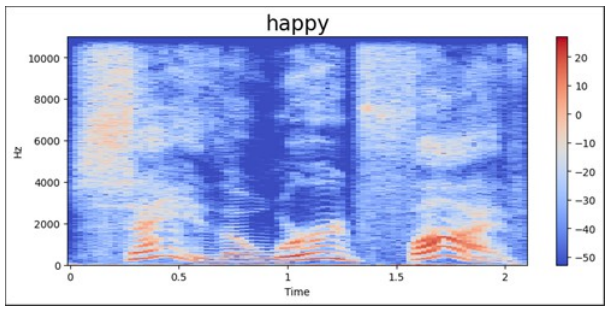


Fig.3.SpectrographoftheAudio

3) *TestingandPerformanceEvaluation:*Thesystemisthen tested with TESS, which provides various emotional speech recordings. The testing process tests the model with various performance metrics such as accuracy, precision, recall, and F1-score to analyze how well the model classifies emotions. Also,aconfusionmatrixisgeneratedtoanalyzetheerrorsand find overlapping emotions, such as fear and surprise, having similar speech patterns. classification, rendering it satisfactory enough for most ap- plications such as mental health analysis, customer service feedback monitoring, and human-computer interaction.
4) *Advantages of the System:*Among the greatest benefits of this system is that it can recognize emotions in real-time with high accuracy. It uses the Wav2Vec2 transformer model, whicheradicatestheneedforhumanefforttofindfeatures, so the system can extract important features from raw speech directly. This improves the accuracy of emotion classification and shortens preparation time. Additionally, through the ad- dition of a Softmax classifier, the system gives accurate and efficient predictions based on probability, and it is simple to identify the dominant emotional states.
   Thesystemisdesignedtobesimpletouse,allowingusersto interact with it easily, and thus many individuals can utilize it. It is also easily expandable since it can be deployed on cloud platforms, allowing it to handle large emotion recognition tasks efficiently. It is not only capable of detecting emotions but can be applied to sentiment analysis, healthcare, virtual assistants, and AI-driven customer support. The capacity to recognize emotions in speech enables the development of morepersonalizedandresponsiveexperiencesforusersacross various industries.
5) *Problems and Limitations:* Despite its advantages, the SER system also has some disadvantages and limitations.One of the principal problems is that it is overly sensitive to background noise, which can alter how well it can classify emotions. Although preprocessing methods can reduce the noise, real environments with sudden sounds can still damage how effective the model is. Moreover, the similarity among certain emotions, such as fear and surprise, makes it difficult to categorize them. They tend to have similar sound and tone patterns, which can confuse the classification process. The computing power required by deep learning models is also challenging,asprocessinginrealtimedemandsaconsiderable amount of power, particularly in the use of edge devices. Enhancing the model for speed of response and maintaining accuracy is an ongoing focus of development. Overcomingthe challenges through changes in data, transfer learning, and enhancing model methods can be used to increase system performance.

VI. RESULTS

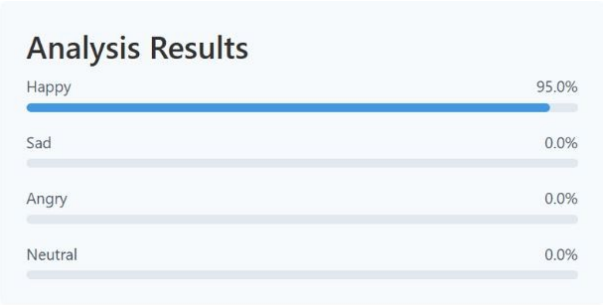


Fig.4.ResultofanAudioInput

The model performs extremely well for highly controlled environments where the speech recordings have minimal orno background noise. Real-world testing reveals noisy condi- tions can suffer a slight reduction in classification accuracy, requiring additional tweaking and methods to address noise. Asageneralprinciple,thesystemprovidesrobustemotion
The Speech Emotion Recognition (SER) system is very accurate, with 95% accuracy and a weighted F1-score of 0.93, andisveryeffectiveinclassifyingemotionsfromspeech.With Wav2Vec2 for feature extraction and a Softmax classifier for classification, the system correctly recognizes neutral, happy, sad, angry, fearful, disgusted, and surprised emotions with high precision and recall. The model is very accurate for detecting subtle emotional differences, particularly for neutral and angry expressions, for confident performance in real-time applications. The system provides real-time emotion predictionswithconfidencescores,anditisveryappropriate formentalhealthscreening,customersentimenttracking, AI-based virtual assistants, and interactive human-computer applications. With its robust deep learning architecture and excellent accuracy, this system can revolutionize emotion- aware AI systems by enabling more intuitive, responsive, and emotionally intelligent interactions in healthcare, customer service, entertainment, and robotics applications.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

VII. CONCLUSION

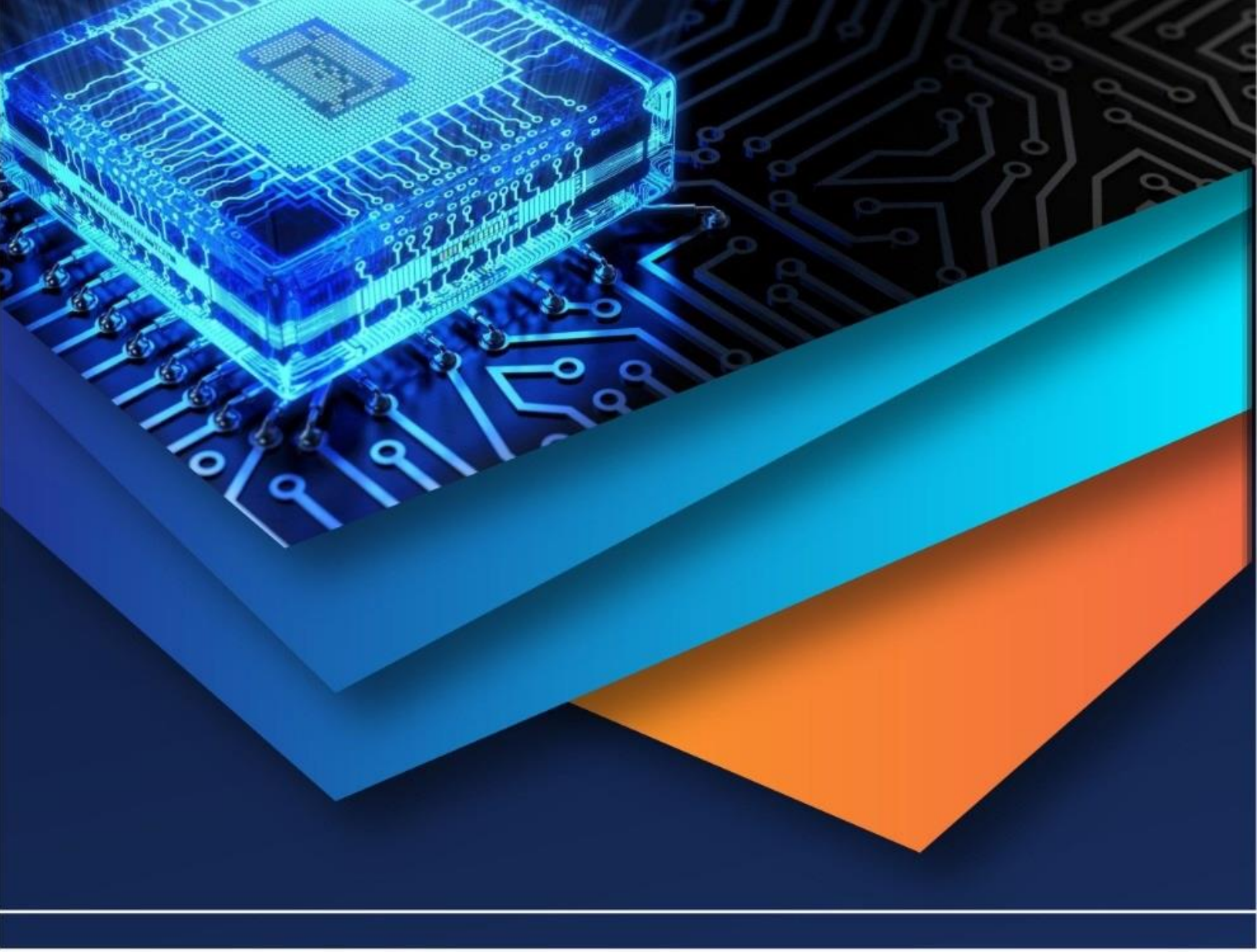
The Speech Emotion Recognition (SER) system presented inthisreporteffectivelyidentifiesemotionsfromspeechusing a Wav2Vec2-based feature extraction and Softmax classifier. Usingdeepspeechembeddings,thesystemaccuratelycaptures phonetic and prosodic variations, enabling accurate emotion identification. The model was trained and tested using bench-markingdatasetswithahighaccuracyof0.95(95%),ensuring its reliability and robustness for emotion classification into neutral, happy, sad, angry, fearful, disgusted, and surprised categories. Through thorough testing and analysis, the system has become a highly efficient and scalable real-time emotion recognition tool. The usage of deep-learning-based methods removestheneedformanualfeatureengineering,making thesystemversatileandapplicabletovariedapplications like mental health monitoring, sentiment analysis, and AI- driven virtual assistants. With future expansion, this system can revolutionize human-computer interaction, promoting a more empathetic and adaptable AI-driven experience across sectors.

VIII. ACKNOWLEDGMENT

Success of a project like this involving high technical expertise, patience, and massive support of guides, is possible when team members work to-gather. We take this opportunity to express our gratitude to those who have been instrumental in the successful completion of this project. We would like to appreciatetheconstantinterestandsupportofourmentorMrs. RohiniDeshpandeinour projectandaidingusindevelopinga flairforthefieldofApplication.Wewouldalwayscherishthe journey of transforming the idea of our project into reality.We would like to show our appreciation to Mrs. Rohini Deshpande for her tremendous support and help, without whom this project would have reached nowhere. We would also like to thank our project coordinator. Dr Chhaya Pawar for providing us with regular inputs about documentation and project timeline. A big thanks to our HOD Dr.M Kiruthikafor all the encouragement given to our team. We would also like to thank our principal, Dr. S.M.Khot, and our college, Fr.C.Rodrigues Institute of Technology, Vashi, for giving us the opportunity and the environment to learn and grow.

REFERENCES

[1] H. I. Attar, N. K. Kadole, O. G. Karanjekar, D. R. Nagarkar, and S.More,"SpeechEmotionRecognitionSystemUsingMachineLearning,"2023.
[2] G. Deshmukh, A. Gaonkar, G. Golwalkar, and S. Kulkarni, "Speechbased Emotion Recognition using Machine Learning," 2023.
[3] G.S.,"SpeechEmotionRecognitionusingMachineLearninginPython," 2023.
[4] D.Bertero,F.B.Siddique,C.-S.Wu,Y.Wan,R.H.Y.Chan,and
[5] P.Fung,"Real-TimeSpeechEmotionandSentimentRecognitionforInteractive," 2023.
[6] T.M.Wani,T.S.Gunawan,andS.A.A.Qadri,"AComprehensiveReview of Speech Emotion Recognition Systems," 2023.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)