Speech Emotion Recognition for Emo-DB Database

BTECH PROJECT REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE

OF

BACHELOR OF TECHNOLOGY IN

ELECTRONICS & COMMUNICATION ENGINEERING

Submitted by:

NIKHIL (2K20/EC/132)

ROHAN CHOUDHARY (2K20/EC/168)

PRITHVIJIT RAKSHIT (2K20/EC/150)

Under the supervision of

DR.SACHIN TARAN

DEPT. OF ELECTRONICS & COMMUNICATION ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

MAY 2024

DEPT. OF ELECTRONICS & COMMUNICATION ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

We, NIKHIL (2K20/EC/132), ROHAN CHOUDHARY (2K20/EC/168) and PRITHVIJIT RAKSHIT (2K20/EC/150) students of B.Tech (Electronics and Communication Engineering), hereby declare that the Project Dissertation titled — "SPEECH EMOTION RECOGNITION USING EMO-DB" which is submitted by us to the Department of Electronics and Communication Engineering, Delhi Technological University, Delhi in fulfillment of the requirement for awarding of the Bachelor of Technology degree, is not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma, Fellowship, or other similar title or recognition.

Place: New Delhi
Date: 30/05/2024

NIKHIL (2K20/EC/132)

ROHAN CHOUDHARY (2K20/EC/168)

PRTITHVIJIT RAKSHIT

(2K20/EC/150)

DEPT. OF ELECTRONICS & COMMUNICATION ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Thesis titled "SPEECH EMOTION RECOGNITION USING EMO-DB" which is submitted by NIKHIL (2K20/EC/132), ROHAN CHOUDHARY (2K20/EC/168) and PRITHVIJIT RAKSHIT (2K20/EC/150) for fulfilment of the requirements for awarding of the degree of Bachelor of Technology (B.Tech) is a record of the project work carried out by the students under my guidance & supervision. To the best of my knowledge, this work has not been submitted in any part or fulfillment for any Degree or Diploma to this University or elsewhere.

Place: New Delhi

Date: 30/05/2024

Dr. SACHIN TARAN
SUPERVISOR
Professor (ECE)
Delhi Technological University

DEPT. OF ELECTRONICS & COMMUNICATION ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

We are extremely grateful to our project guide, Dr. Sachin Taran, Assistant Professor, Department of Electronics and Communication Engineering, Delhi Technological University, Delhi for providing invaluable guidance and being a constant source of inspiration throughout our research work. We will always be grateful to him for his extensive support and encouragement.

We are extremely grateful to all the panel members who evaluated our progress, guided us throughout our project, and gave us constant support and motivation, innovative ideas, and all the information that we needed to pursue this project idea.

We would like to extend our vote of thanks to all our college mates and acquaintances who provided strong emotional and familiar support to keep us motivated.

Thank You!

ROHAN CHOUDHARY	NIKHIL	PRITHVIJIT RAKSHIT
(2K20/EC/168)	(2K20/EC/132)	(2K20/EC/150)

ABSTRACT

Keywords – Emotion Recognition , Silence Removal , Feature Extraction , KNN classifier

Speech is a foundational and primary form of human communication, allowing distinct individuals to convey thoughts and notions for their community and own ideas, and information to others. This study achieved 91.8% accuracy in identifying the emotions in German speech same method can be performed in datasets of other languages.

The study focuses on a proposed structure for performing emotion recognition in German speech, which includes silence removal, feature extraction of speech, feature selection, and feature classification. The process first includes preprocessing, in which silence in the speech is eliminated using short-time energy and zero crossing rate.

For the study only, the Emo-DB database which stands for “Emotional Speech Berlin Database” will be used. The proposed structure aims to provide the highest accuracy with the KNN classifier, as evaluated against other conventional deep-learning methods.

CONTENTS

List of Tables

List of Figures

LIST OF SYMBOLS, ABBREVIATIONS AND NOMENCLATURE

SER – Speech Emotion Recognition

CNN – Convolutional Neural Network

KNN – K nearest Neighbours

FFT – Fast Fourier Transform

MFCC – Mel-frequency cepstral coefficient

ZCR – Zero Crossing Rate

GTCC – gamma tone cepstral coefficient

EMO-DB – Emotional Database

INTRODUCTION

Speech is the quickest, most efficient, and most practical means of human-to-human communication and conveying information, it is the main stakeholder in the communication paradigm. Speech is an intricate signal with a wealth of valuable information, and studying the emotional nuances of the speech signal is the primary goal of this study. Speech Signal is a mix of complex frequencies resulting in a complex wave [1]. The difficult procedures in the analysis of the speech signal are a fascinating challenge in the field of speech emotion recognition (SER), which has attracted a lot of interest in recent research around speech emotion recognition. SER is the method of extraction and analysis of emotional aspects of a speech signal, it includes verbal entities and some prosodic entities like emphasis, intonation, and rhythm. The Emotional dynamics can be properly understood by studying the minor fluctuations in intensity, arousal, valence, and other characteristics which add to the intricacy of an emotion. In this field of study, researchers work towards building strong computational models that can reliably identify and classify emotional states from voice data [1-2].

Classical Speech Emotion Recognition systems focus less on the variability of the emotion parameters in speech induced through gender, state of mind, culture, etc. Emotion Recognition has been an exploration topic for a much longer time, The most basic form of emotion recognition systems includes detecting emotion from facial expressions, although that includes several challenges in detecting emotion through this form of emotion extraction. The base of Human-computer Interaction Systems is controlled through emotions, in the contemporary era of this world, speech emotion recognition (SER) is acquiring popularity among scholars as the nuances in emotion extraction are nevertheless challenging from human speech [3].

One of the significant hurdles and major Challenges in SER is the need to identify the set of emotions to be targeted from raw speech signals, language experts have crafted lists of the emotional conditions most commonly encountered in our daily experiences. This problem is also known as emotions classification from speech, classification of large emotion sets is immensely difficult. Researchers have agreed with the ‘palette theory’ which says that any emotion can be classified and decomposed into sets of primary emotions similar to the manner such that any color results from a mixture of fundamental hues. Emotion and feature classification is one of the major issues in recent research that is considered in the process of ser that this study has focused on solving [4].

Several methods are used for understanding the fundamental emotional constituents of a speech signal but the majority of difficulty in the analysis of emotions from speech falls back on the processing of the raw speech signal and on the addition of noise and variable parameters in a speech the complexity grows even further, several types of research have aimed to tackle the lying problem through various deep learning and machine learning models but there is a gap in an effective solution and actual problem therefore, there is space for much more advancements in this area of research [5]. Most of the approaches that have been used to analyze human speech in the context of emotion recognition had some shortcomings, Fast-Fourier transform (FFT) had an issue of loss of information in conversion [5-6]. FFT drawbacks were tackled by short-time Fourier transform, but still, the stationary and non-stationary behavior of the speech signal was a matter of concern [6].

Existing drawbacks in SER include loss of signal in the extraction of speech data from the raw signal, this study is targeted towards minimizing the loss using multi-layered architecture to treat the raw speech signal from the Emo-DB dataset for the analysis of emotional constituents [7-8]. Algorithmic analysis of speech comes with the drawback of losing originality in the raw speech data, therefore feature extraction and classification have taken the attention of scholars in recent studies. This study focuses around minimizing the loss of originality in the raw speech data and generating better emotional cues from a monolingual speech. Using this targeted method of analyzing and pre-processing the data, this study has achieved an accuracy of 91.8% in the extraction of emotional constituents from a monolingual speech signal [8]. The feature Classification technique used in this study revolves around the K-Nearest Neighbors (KNN) classifier for speech classification. KNN is a simple and efficient algorithm used for refining emotional states from speech data, thus is widely used in speech emotion recognition techniques [8-9]. The Choice of the KNN algorithm has allowed us to go deeper into the emotional parameters in a speech signal, allowing high precision in gathering data from the German Speech. The sequence of methods followed has been a major factor in achieving accuracy and thus moving forward with the study of speech emotion recognition [9]. The following speech processing architecture has been followed throughout this study: Emo-Db Dataset, Silence Removal, Feature Extraction, Reconstruction, Feature Selection, Feature Classification, and Seven Class SER [9-10]. Every step followed in this architecture from the very beginning of the raw speech signal from Emo-DB and pre-processing it through various noise removal processes and then the reconstruction and classification of the signal is crucial in extracting emotional constituents from speech.

Understanding human speech and extracting emotions is of the utmost need in today's time, most of the practical implementations of speech emotion recognition would be in the healthcare industry where the psychological state could be an indicator of the patient's mental state and this could be identified by using emotional status through the speech of the patients [10]. Another use-case of Ser is in the Law-Justice system where emotional analysis could be used to detect the truth in a speaker's actual statement from reality [10-11]. These SER Systems help in tackling various such problems that exist in human society and are difficult to sustain from a human perspective, there are the areas where speech emotion recognition is of most use. Emotions have always been a difficult topic for research due to the lack of understanding of the variable nature of the human mind [11]. There are many existing ways of understanding emotion but from an algorithmic perspective, analysis of speech takes over a bit more priority than traditional methods, due to the constraints in other methods. Recording and analyzing speech data is more feasible for datasets algorithms have proven to be more efficient and accurate than other ways.

SELECTING OF SPEECH DATASET

Selecting a speech dataset for speech emotion recognition (SER) research involves careful consideration of several key factors to ensure the dataset is suitable for your research objectives and methodologies. Here's a comprehensive guide on how to select an appropriate speech dataset for SER:

1. Define Research Objectives

Scope of Emotions: Determine the range of emotions you aim to recognize (e.g., basic emotions like happiness, sadness, anger, or more nuanced states like frustration or calmness).

Application Context: Identify the specific application or context for the SER system (e.g., customer service, healthcare, human-computer interaction).

2. Dataset Characteristics

Emotion Labels: Ensure the dataset includes well-defined and annotated emotion labels. These labels should be consistent and validated by multiple annotators to ensure reliability.

Data Diversity: Look for diversity in speakers (gender, age, accent), languages, and recording conditions to enhance the generalizability of your model.

Audio Quality: Check the audio quality, including sampling rate and presence of background noise. Higher quality recordings can improve model performance but may not represent real-world conditions.

3. Size and Balance

Dataset Size: A larger dataset can provide more examples for training and testing, leading to better model performance. Ensure the dataset has enough samples per emotion class to train a robust model.

Class Balance: Verify that the dataset has a balanced distribution of emotion classes. Imbalanced datasets can lead to biased models.

4. Metadata and Annotations

Detailed Metadata: Look for datasets with rich metadata, including speaker demographics, recording conditions, and textual transcriptions if available.

Annotation Quality: High-quality annotations are crucial. Check if the annotations were done by experts and if there are multiple annotators for cross-validation.

5. Licensing and Availability

License Type: Ensure the dataset is available for research use and complies with your institution's and funder's data usage policies. Open-source datasets are typically preferred for academic research.

Accessibility: The dataset should be easily accessible without overly restrictive conditions. Consider datasets that are well-documented and supported by the research community.

6. Popular Speech Emotion Recognition Datasets

Consider some widely used and reputable datasets in the SER field:

IEMOCAP (Interactive Emotional Dyadic Motion Capture Database): Contains acted emotional dialogues with multimodal data (audio, video, text). It's extensively used in SER research.

CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset): Contains audio-visual recordings of actors portraying a range of emotions. Useful for multimodal SER.

RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song): Offers both emotional speech and song, recorded by professional actors.

MSP-IMPROV (Multimodal SPontaneous Improvisation): Includes natural, spontaneous dialogues designed to elicit emotional responses.

7. Evaluation and Benchmarking

Baseline Models: Check if there are existing baseline models or benchmarks for the dataset. This can help you evaluate the performance of your SER models and compare with state-of-the-art methods.

Community Usage: Prefer datasets that are widely used in the community, as this often means better support, more resources (like preprocessed data or additional tools), and established benchmarks.

SPEECH EMOTION DATASET

In this study, the publicly available dataset Emo-DB is used, short for the "Berlin Database of Emotional Speech," containing voice audio recordings of acted emotional speech in the German language. It is extensively used in research related to speech emotion recognition, affective computing, and related fields.

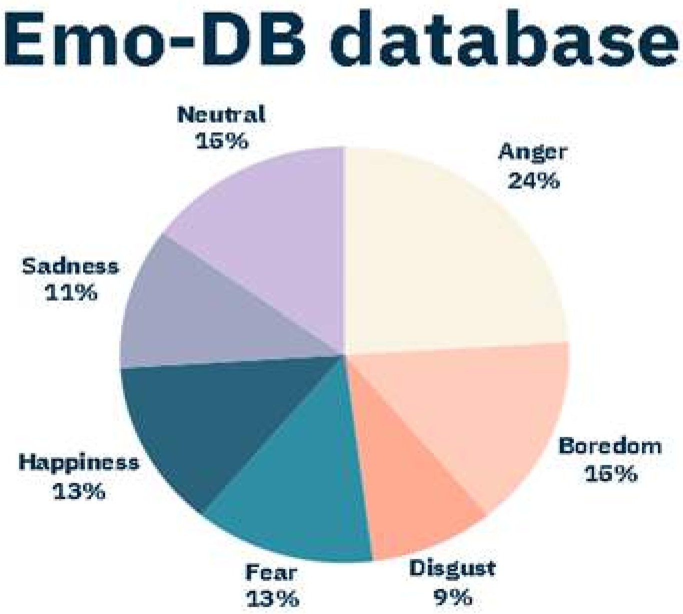


Fig 1 Percentage distribution of emotions in the Emo-DB database

A. Emo-DB

This German dataset has a voice of a total of 10 individuals (5 males and 5 females) all the speakers are professional individuals, this data contains 535 expressions and consists of seven emotional states of humans which are Boredom, anger, happiness, anxiety, neutral, disgust, and sadness. Data is available in 16khz after down-sampling but it was initially recorded at a 48 kHz sampling rate [12]. The total percentage of each type of speech in humans is represented in Fig.1.

METHODOLOGY

The process in this paper involves the following steps, initially, the silent part of the speech is removed using zero crossing rate and short-term energy step by step. After this step, feature extraction, feature selection, and feature classification are performed using MFCC, relief algorithm, and KNN classifier respectively after all these steps results are generated in MATLAB and further analyzed to obtain the result. With the integration of these advanced speech processing steps, the proposed methodology accurately identifies emotions in the German language and gives the highest accuracy in comparison to any other available method that uses a KNN classifier. The block diagram shown in Fig. 2 describes the flow of the process for each feature used in this study.

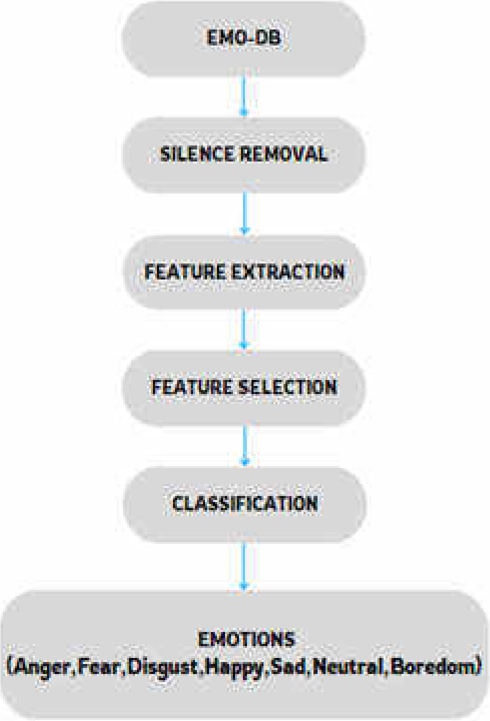


Fig.2 Flow diagram of the process used

A. Removing silence from the dataset

Removing silenced audio from the signal is one of the main tasks in this process. Silence removal serves as a critical preprocessing step, essential for refining audio data before subsequent analyses and applications. This involves removing the silent part or say, a signal with absolutely zero amplitude from the part of speech that is meaningful for the process. Techniques for silence removal typically employ sophisticated signal processing methods focused on carefully and accurately removing the no amplitude region while maintaining the authenticity and usefulness of other speech content [13]. Inspection of speech signal data is carried out using short-time audio signal processing methods like short-time Fourier transform (STFT) which examines signal in small frames. In Eq.1 short-term Fourier transform is represented, short-term energy(E) is represented in Eq.2, the Zero-crossing rate (ZCR) is represented in Eq.3, and finally Entropy of the energy or speech signal which is represented through Eq.4, all equations are taken from the reference [14]. Raw speech signal from the Emo-DB dataset and the same signal after removing the silenced part from it can be seen in Fig.3. This refined data can then be used for audio summarization, music visualization, or dimensionality reduction. In the end, silence removal from the audio signal helps significantly increase the output of the overall result and minimize the chance of inaccuracy also the utility of the subsequent audio signal increases significantly.

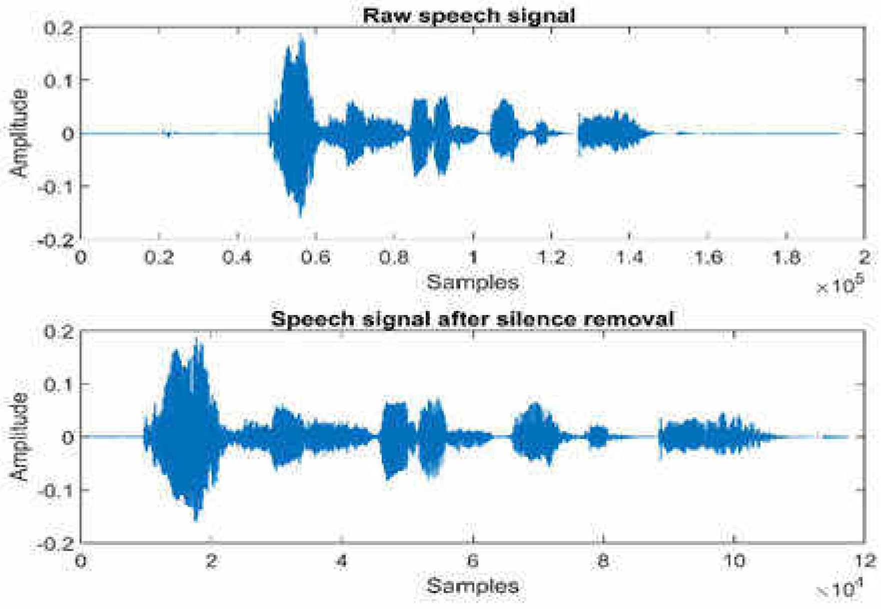


Fig.3 Speech signal before and after silence removal

$$X(t,f) = \int_{-\infty}^{\infty} x(\tau)w(\tau - t)e^{-j2\pi f\tau}d\tau \qquad (1)$$

$$E = \sum_{n=0}^{N-1} |x(n)|^2 \qquad (2)$$

$$ZCR = \frac{1}{N-1}\sum_{n=1}^{N-1} |\text{sgn}\,(x(n)) - \text{sgn}\,(x(n-1))| \qquad (3)$$

$$H = -\sum_{i=1}^{N} p_i \log_2\,(p_i) \qquad (4)$$

where N, n, t, and f denote the length of the frame, number of samples, time, and frequency of speech signal respectively.

B. Feature extraction

In this paper, periodic and structural features of speech are extracted from each frame with a duration of 80 milliseconds with 75% overlapping with the adjacent frame. There are a total of 105 features are computed from each frame of speech signal using the feature extraction algorithm and after that 50 potential features are selected for classification purposes. After the extraction of potential features edge smoothening of the frame is necessary which is done by hamming window [15-19], now from each frame gamma tone cepstral coefficient (GTCC), Mel-frequency cepstral coefficient (MFCC), pitch, spectral crest, spectral entropy, spectral kurtosis, spectral skewness, harmonic ratio, and spectral centroid are computed. Subsequently, the extracted features are utilized in the classification of German language speech emotions, harnessing the discriminative power of these features to differentiate between different emotional states expressed in speech [20]. This methodology emphasizes or is more based on the importance of feature extraction in capturing salient characteristics of speech signals and highlights its pivotal role in subsequent analysis tasks such as emotion classification.

C. Feature selection

In this study, for the feature selection of speech procedure utilize the Relief algorithm as a specialized feature selection method tailored to the task of emotion identification in speech, particularly focusing on the Emo-DB database. The Relief algorithm works by iteratively evaluating the discriminative power of each acoustic speech feature extracted from speech signals, the mathematical representation for the Relief algorithm to determine the feature score represented in Eq.5 from the reference [21].

This feature compares the differences in feature values between instances belonging to the same and different speech emotion classes, the algorithm assigns significance to each feature, which reflects its ability to between emotions. Using this targeted approach, this study aims to identify a subset of acoustic features that demonstrate the highest discriminative power for emotion identification in speech, thereby enhancing the accuracy and efficiency of this emotion recognition method while minimizing computational complexity.

$$
\begin{aligned}
f_i = & \frac{1}{c}\sum_{q=1}^{R}\left(-\frac{1}{m_q}\sum_{x_j \in NS(q)} d(X(q,i)-X(l,i))\right) \\
& \square + \sum_{z \neq z_q} \frac{1}{h_{qz}} \frac{p(z)}{1-p(z)} \sum_{x_j \in NH(q,z)} d(X(q,i)-X(l,i))
\end{aligned}
\qquad (5)
$$

In this equation number of classes is represented by c, NH(q,y), and NS(q) are the closest points from xq in the class z which have size hqz and xq respectively, and p(y) represents the ratio of instances for class z.

D. Feature Classification

K -Nearest Neighbor (KNN) is used as a speech emotion feature classifier in this study, it is chosen because of its simplicity and effectiveness in analyzing and discriminating emotional states from speech signals using Euclidean distance computation which is mathematically represented in Eq.6 and taken from the reference [22]. Speech signal acoustic features are extracted and selected after that KNN classifier algorithm computes the distance between consecutive and non-consecutive test samples of the speech and the training samples, identifying the K nearest neighbors and assigning the majority class to determine the emotional states of the speech. it enhances the precision in capturing subtle emotional nuances.

This technical approach highlights KNN's efficacy in advancing speech-emotion recognition systems [23-25].

DETAILED ANALYSIS OF EMOTIONS IN SPEECH EMOTION RECOGNITION (SER)

Speech Emotion Recognition (SER) focuses on identifying human emotions through vocal expressions. Recognizing specific emotions such as anger, boredom, disgust, fear, happiness, neutral, and sadness is crucial for enhancing the efficacy of SER systems. Here is an in-depth analysis of each emotion’s significance in the context of SER:

1)Anger: Significance in SER:

Acoustic Features: Anger often manifests through increased pitch, intensity, and tempo. The voice might sound louder and sharper.

Applications: Recognizing anger can be crucial in customer service to handle dissatisfied customers promptly, in conflict resolution scenarios, and for monitoring aggression in mental health contexts.

2)Boredom :Significance in SER:

Acoustic Features: Boredom is typically characterized by a low pitch, reduced intensity, and slower speech rate. The voice might sound monotonous.

Applications: Detecting boredom is useful in educational settings to adjust teaching methods, in entertainment to gauge audience engagement, and in workplace productivity monitoring.

3)DisgusT: Significance in SER:

Acoustic Features: Disgust can be identified by a nasal tone, lower pitch, and sometimes a slower speech rate. It may also include sounds of revulsion.

Applications: In healthcare, recognizing disgust can help in diagnosing psychological conditions. It's also valuable in content moderation and improving human-computer interaction.

4)Fear: Significance in SER:

Acoustic Features: Fear is associated with higher pitch, faster speech rate, and a trembling or quivering voice. There may also be irregular breathing patterns.

Applications: Identifying fear is critical in security and emergency response systems, mental health monitoring, and virtual assistants to provide appropriate support during stressful situations.

5)Happiness: Significance in SER: Acoustic Features: Happiness is often reflected through higher pitch, greater intensity, faster speech rate, and a brighter, more dynamic tone.

Applications: Recognizing happiness can enhance user experience in interactive systems, monitor social dynamics in team settings, and improve customer satisfaction analysis.

6) Neutral: Significance in SER:

Acoustic Features: Neutral speech usually has a moderate pitch, consistent speech rate, and steady intensity. It lacks the distinct variations found in emotional speech.

Applications: Detecting neutrality serves as a baseline to distinguish between other emotions. It's essential in everyday interactions and helps in calibrating SER systems to detect deviations towards more emotional states.

7) Sadness: Significance in SER: Acoustic Features: Sadness is often conveyed through lower pitch, slower speech rate, reduced intensity, and a more monotonous tone. The voice might sound soft and low-energy.

Applications: Recognizing sadness is important in mental health monitoring, providing support in customer service scenarios, and in enhancing empathetic responses in virtual assistants and robots.

CURRENT APPLICATION OF SER

1. Customer Service and Call Centers

Emotion Detection: SER can detect customer emotions during calls, helping agents respond more empathetically and effectively.

Real-Time Analytics: Provides real-time feedback to supervisors about the emotional states of both customers and agents, allowing for immediate intervention if necessary.

Personalized Service: Helps tailor responses and solutions based on the emotional state of the customer, improving overall satisfaction.

2. Healthcare and Mental Health

Therapeutic Monitoring: Monitors patients' emotional states during therapy sessions, aiding therapists in understanding their clients' progress and emotional well-being.

Telemedicine: Enhances remote consultations by providing doctors with additional emotional context, which is crucial for accurate diagnosis and treatment.

Mental Health Apps: Integrates with apps designed to monitor and support mental health, providing users with feedback and resources based on detected emotional states.

3. Human-Computer Interaction

Virtual Assistants: Improves the responsiveness and empathy of virtual assistants (like Siri, Alexa, Google Assistant) by enabling them to detect and respond to users' emotions.

Social Robots: Enhances the interaction between humans and robots, making robots more effective in social roles such as companions for the elderly or autistic children.

Interactive Voice Response (IVR) Systems: Enhances automated phone systems by allowing them to adjust their responses based on detected emotions, improving user experience.

4. Education and E-Learning

Adaptive Learning: Adjusts teaching strategies based on the emotional states of students, helping to maintain engagement and optimize learning outcomes.

Feedback Mechanisms: Provides teachers and instructors with insights into student emotions, allowing for more personalized and effective teaching.

5. Entertainment

Interactive Gaming: Enhances the gaming experience by adjusting game dynamics based on players' emotions, creating more immersive and responsive gameplay.

Content Recommendation: Adjusts content recommendations on streaming platforms based on users' emotional responses to different types of content.

FUTURE APPLICATION OF SER

Technological Advancements Driving SER Forward

The future of Speech Emotion Recognition (SER) will be shaped by several technological advancements and trends, which will enhance its capabilities and broaden its applications. Here are some key areas of development:

1. Improved Machine Learning Algorithms

Deep Learning: Leveraging deep neural networks to improve the accuracy and robustness of SER models. Deep learning can capture complex patterns in speech that traditional methods may miss.

Transfer Learning: Using pre-trained models to transfer knowledge from one domain to another, reducing the amount of labeled data required and improving performance in low-resource settings.

Reinforcement Learning: Implementing reinforcement learning to adapt SER systems dynamically based on user interactions and feedback, making them more responsive over time.

2. Multimodal Emotion Recognition

Combining Modalities: Integrating speech with other modalities such as facial expressions, body language, and physiological signals (e.g., heart rate, galvanic skin response) for a more comprehensive understanding of emotions.

Context-Aware Systems: Developing systems that consider the context of the interaction (e.g., environment, topic of conversation) to improve the accuracy of emotion detection.

3. Real-Time Processing

Edge Computing: Implementing SER on edge devices to enable real-time emotion detection without the need for cloud processing, improving response times and preserving privacy.

Optimized Algorithms: Creating more efficient algorithms that can process speech data quickly and with minimal computational resources, making SER feasible for a wide range of devices.

4. Natural Language Processing (NLP) Integration

Sentiment Analysis: Combining SER with sentiment analysis to provide a more nuanced understanding of users' emotional states by analyzing both the content and tone of speech.

Conversational Agents: Enhancing conversational agents with SER capabilities to enable more natural and emotionally aware interactions.

ETHICAL AND THD SOCIAL IMPLICATION

As SER technology advances, it will be crucial to address its ethical and social implications:

1. Privacy and Consent

Informed Consent: Ensuring that users are fully informed about how their emotional data will be used and obtaining their explicit consent.

Data Security: Implementing robust security measures to protect emotional data from unauthorized access and breaches.

2. Bias and Fairness

Diverse Datasets: Using diverse datasets to train SER systems to minimize biases and ensure fair treatment across different demographic groups.

Continuous Monitoring: Regularly auditing SER systems for bias and updating models to mitigate any detected biases.

3. Psychological Impact

User Autonomy: Respecting user autonomy by allowing individuals to control how their emotional data is used and by providing options to opt-out.

Impact on Human Interaction: Studying the impact of SER on human interactions and ensuring that it enhances rather than detracts from genuine human connection.

VISION FOR THE FUTURE

The vision for the future of SER encompasses a world where technology seamlessly understands and responds to human emotions, improving the quality of interactions and providing support across various domains. Here are some futuristic scenarios where SER could play a pivotal role:

1. Smart Cities

Public Safety: Utilizing SER in public safety systems to detect stress or distress in real-time, enabling quicker responses to emergencies.

Public Services: Enhancing public service interactions by tailoring responses based on citizens' emotional states, improving satisfaction and engagement.

2. Personalized Healthcare

Home Monitoring: Integrating SER into home health monitoring systems to provide continuous assessment of patients' emotional well-being and alert caregivers to any concerning changes.

Proactive Interventions: Developing proactive intervention strategies based on emotional state detection, helping to prevent crises before they occur.

3. Educational Transformation

Emotionally Intelligent Tutoring Systems: Creating intelligent tutoring systems that adapt to students' emotions, providing personalized support and encouragement to optimize learning outcomes.

Inclusive Education: Using SER to better support students with special needs by understanding and responding to their emotional cues, fostering an inclusive learning environment.

4. Advanced Human-Robot Interaction Empathetic Robots: Designing robots that can understand and respond to human emotions, providing companionship and support in settings like eldercare, childcare, and therapy.

Collaborative Workspaces: Implementing SER in collaborative robots (cobots) to enhance teamwork between humans and robots in various industries, improving productivity and safety.

Importance in SER: The F1 score provides a single metric that balances both the precision and recall of your SER system. This is particularly valuable when you have an uneven class distribution.

Handling Imbalance: Emotions in speech datasets can be imbalanced (e.g., neutral speech may be more common than anger), and the F1 score helps provide a balanced measure of performance, highlighting models that perform well across both common and rare emotions.

Specific Scenarios in SER

Customer Service: High precision is needed to avoid false alarms, and high recall ensures all instances of customer dissatisfaction are detected.

Mental Health Monitoring: High recall is critical to ensure all signs of distress are detected, while precision ensures that false positives do not lead to unnecessary interventions.

Human-Robot Interaction: Both precision and recall are vital to maintain trust and reliability in interactions, ensuring the system responds appropriately to the user's emotional state.

RESULT AND DISCUSSION

The study described here delves into emotion recognition using the Emo-DB database and a systematic processing approach outlined in the paper. The first step for the process of SER silenced part is removed from the speech using ZCR and short-term energy, after eliminating silent segments from speech signals, the methodology involves feature extraction, in which every utterance of the speech signal is divided into frames of length

80 milliseconds each with 75% overlapping as explained in section. III. Further only 50 features are selected for the KNN classifier from a total of 105 features of each frame using a hamming window after this using the Relief algorithm differences are calculated between feature values for the feature selection of the speech signal, in the last study used the KNN classifier to compute Euclidean distance for the classification of the features. The achieved overall accuracy on the Emo-DB dataset is an impressive 91.8%, showcased in the final results, and further F1 score, precision, and recall are calculated for each speech emotion through the confusion matrix in Fig.4, feature rank bar for the dataset which is shown in Fig.5, majorly helps to identify visually which feature contributes most significantly to build the result or outcome of the whole procedure this also helps in identifying relevant variables of feature for the classification purpose. On the other side, the ROC curve which is represented in Fig.6 shows how well the model is performing in identifying positive and negative classes, it shows how well the classifier is good at discriminating between the subsequent classes of signal data.

In Table I. precision, recall, and F1 score is shown for all seven emotions in the Emo-DB dataset it is evident from the

table that the precision value for anger is highest at 95.82% followed by happy at 92.72% and disgust at 91.50%. The recall value for anger 94.33% is the highest and this time followed by neutral 92.60% and disgust 92.37% finally the F1 score is highest for anger 95.07% which represents how much this methodology is identifying instances of specific emotions and minimizing the misclassification of the emotion. This study's main key points are in recognizing the speech features for accurate emotion recognition, for the main or notable emotions like anger, happiness, disgust, and other significant emotions in human life.

The result shown in Table I clearly shows that the F1 score, precision, and recall are highest in the case of anger emotion compared to other emotions, which suggests that the proposed methodology has the highest sensitivity for anger. Emotions like sadness, boredom, and fear have less precession and recall values compared to others in which sadness has the lowest F1 score.

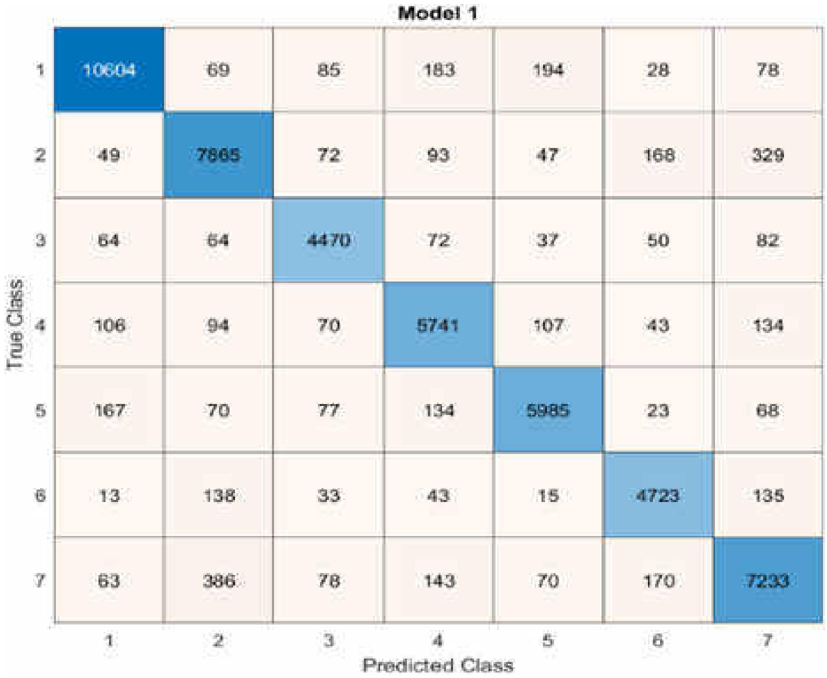
In table II. approaches for SER and their accuracy for Emo-DB dataset is compared with each other study in the reference [25] focus more on the effectiveness of feature selection techniques and classifier models to obtain high accuracy of 91.16% using RNN and 86.22% using SVM. on the other hand, study in paper [26] proposes a CNN based feature extraction approach which is combined with spectrograms for achieving the accuracy of 85.57% for Emo-DB database, study in reference [27] introduced a feature selection procedure which is mainly based on emotional difference achieving overall accuracy of 84.62%. further research shows the superiority of Mel frequency magnitude coefficient for emotions recognition of different database showing 81.5% accuracy [28], and finally methodology in reference [29] goes for hybrid approach combining twine-shufpat, INCA and TQWT techniques resulting in achieving accuracy of 79.08% for Emo-DB database.

Table I. Result for Emo-DB proposed methodology

Emotions	F1 score (%)	Precision (%)	Recall (%)
Anger	95.07	95.82	94.33
Boredom	90.87	90.54	91.21
Disgust	91.93	91.50	92.37
Fear	90.34	89.57	91.12
Happy	92.22	92.72	91.73
Neutral	92.67	90.74	92.60
Sad	89.28	89.75	88.62

Table II. Result comparison of studies for the Emo-DB

Author	Classifier	Accuracy (%)
Kerkeni [26]	SVM	86.66
Mustaqeem [27]	RBFN	85.57
O¨zseven [28]	SVM	84.62
Ancilin [29]	MSVM	81.5
Tuncer [30]	SVM	79.08
Proposed methodology	KNN	91.8



	1	2	3	4	5	6	7
1	10604	69	85	183	194	28	78
2	49	7865	72	93	47	168	329
3	64	64	4470	72	37	50	82
4	106	94	70	5741	107	43	134
5	167	70	77	134	5985	23	68
6	13	138	33	43	15	4723	135
7	63	386	78	143	70	170	7233

Fig.4 Confusion matrix

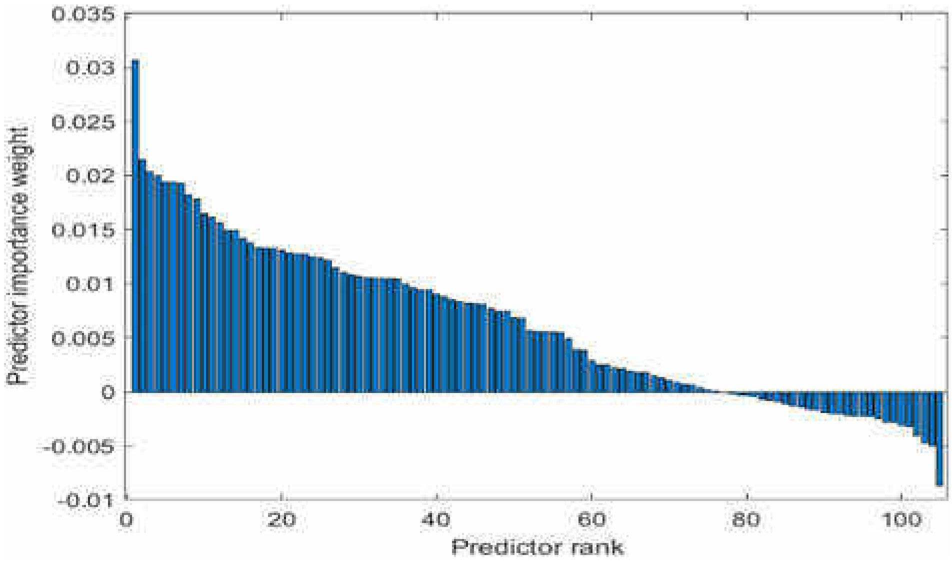


Fig.5 Feature predictor rank bar

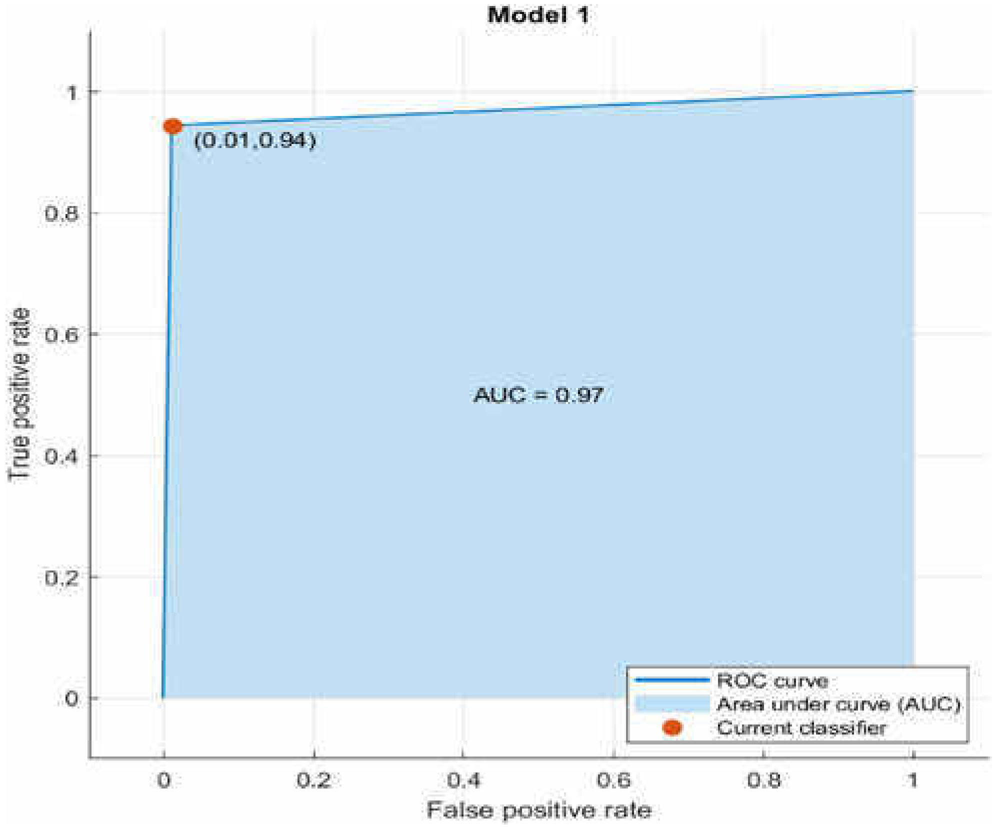


Fig.6 ROC curve of Emo-DB database

CONCLUSION

The proposed framework in this study is an attempt to leverage feature extraction using frames, feature selection by utilizing a relief algorithm, and finally KNN, to classify speech emotions from various datasets. Considering the Emo-DB dataset, specifically, this paper addressed the challenges associated with analyzing emotions in speech signals. Using the methods described in the section. III for preprocessing speech signals this study able to achieve a total accuracy of 91.8 for all classes of the emotion of the Emo-DB dataset. Results of various studies are compared for the German speech dataset (Emo-DB) and it is also shown in the table which classifier that study used, it is very clear from the table that the proposed methodology achieved the highest accuracy among all studies that used KNN only as their classifier. Our study concludes the proposed structure is efficient and flexible in tackling the complexities of various speech signals. But this research is mostly based on the emotions that are acted and it may fully resemble natural human speech emotions perhaps the KNN classifier is effective for this dataset other algorithms and methodology should also be explored to improve the performance of the result. It even can contribute various insights into the fields of speech and emotion recognition in machine learning in further studies. Future research will be conducted on refining existing methodologies and exploring interdisciplinary collaborations to deepen our understanding of emotion recognition through speech signals.

REFERENCES

[1] J. Blackledge, D. Kearney, C. Farrell, and G. Kearney, "Energy Commodities Trading Using a Phase Signal Derived from the Levy Index of Price and Volatility," 2012.

[2] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Information Fusion*, vol. 59, pp. 103-126, 2020.

[3] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A Comprehensive Review of Speech Emotion Recognition Systems," *IEEE Access*, vol. 9, pp. 47795-47814, 2021.

[4] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572-587, 2011.

[5] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 59, p. 101894, 2020.

[6] A. F. Haque, "Frequency analysis and feature extraction of impressive tools," *International Journal of Advance Innovations, Thoughts & Ideas*, vol. 2, no. 2, pp. 1, 2013.

[7] B. Chen, Q. Yin, and P. Guo, "A study of deep belief network based Chinese speech emotion recognition," in *2014 Tenth International Conference on Computational Intelligence and Security*, 2014, pp. 180-184.

[8] M. Nayak and B. S. Panigrahi, "Advanced signal processing techniques for feature extraction in data mining," *International Journal of Computer Applications*, vol. 19, no. 9, pp. 30-37, 2011.

[9] M. A. Jalal, E. Loweimi, R. K. Moore, and T. Hain, "Learning temporal clusters using capsule routing for speech emotion recognition," in *Proceedings of interspeech 2019*, 2019, pp. 1701-1705.

[10] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[11] W. J. Yoon, Y. H. Cho, and K. S. Park, "A study of speech emotion recognition and its application to mobile services," in *Ubiquitous Intelligence and Computing: 4th International Conference, UIC 2007, Hong Kong, China, July 11-13, 2007. Proceedings 4*, 2007, pp. 758-766.

[12] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Interspeech*, vol. 5, 2005, pp. 1517-1520.

[13] S. Taran, "A novel decomposition-based architecture for multilingual speech emotion recognition," Neural Computing and Applications, pp. 1-13, 2024.

[14] T. Giannakopoulos, "A method for silence removal and segmentation of speech signals, implemented in Matlab," University of Athens, Athens, 2, 17, 2009.

[15] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, "Speech emotion recognition: Features and classification models," Digital Signal Processing, vol. 22, no. 6, pp. 1154-1160, 2012.

[16] M. A. Hossan, S. Memon, and M. A. Gregory, "A novel approach for MFCC feature extraction," in 2010 4th International Conference on Signal Processing and Communication Systems, pp. 1-5, Dec. 2010.

[17] Q. Wu, L. Zhang, and G. Shi, "Robust multifactor speech feature extraction based on Gabor analysis," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 927-936, 2010.

[18] N. Sharma, M. H. Kolekar, K. Jha, and Y. Kumar, "EEG and cognitive biomarkers based mild cognitive impairment diagnosis," IRBM, vol. 40, no. 2, pp. 113-121, 2019.

[19] J. Kim and R. A. Saurous, "Emotion Recognition from Human Speech Using Temporal Information and Deep Learning," in Interspeech, pp. 937-940, Sep. 2018.

[20] S. Taran, "Emotion recognition using rational dilation wavelet transform for speech signal," in *2021 7th International conference on signal processing and communication (ICSC)*, 2021, pp. 156-160.

[21] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM computing surveys (CSUR)*, vol. 50, no. 6, pp. 1-45, 2017.

[22] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, "Efficient kNN classification with different numbers of nearest neighbors," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 5, pp. 1774-1785, 2017.

[23] A. Ali, M. A. Alrubei, L. F. M. Hassan, M. A. Al-Ja'afari, and S. H. Abdulwahed, "Diabetes Diagnosis based on KNN," *IIUM Engineering Journal*, vol. 21, no. 1, pp. 175-181, 2020.

[24] A. Bhavan, P. Chauhan, and R. R. Shah, "Bagged support vector machines for emotion recognition from speech," *Knowledge-Based Systems*, vol. 184, p. 104886, 2019.

[25] Z. Soumaya, B. D. Taoufiq, N. Benayad, K. Yunus, and A. Abdelkrim, "The detection of Parkinson disease using the genetic algorithm and SVM classifier," *Applied Acoustics*, vol. 171, p. 107528, 2021.

[26] L. Kerkeni, Y. Serrestou, K. Raoof, M. Mbarki, M. A. Mahjoub, and C. Cleder, "Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO," *Speech Communication*, vol. 114, pp. 22-35, 2019.

[27] M. Mustaqeem, M. Sajjad, and S. Kwon, "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM," *IEEE Access*, vol. 8, pp. 79861-79875, 2020.

[28] T. Özseven, "A novel feature selection method for speech emotion recognition," *Applied Acoustics*, vol. 146, pp. 320-326, 2019.

[29] J. Ancilin and A. Milton, "Improved speech emotion recognition with Mel frequency magnitude coefficient," *Applied Acoustics*, vol. 179, p. 108046, 2021.

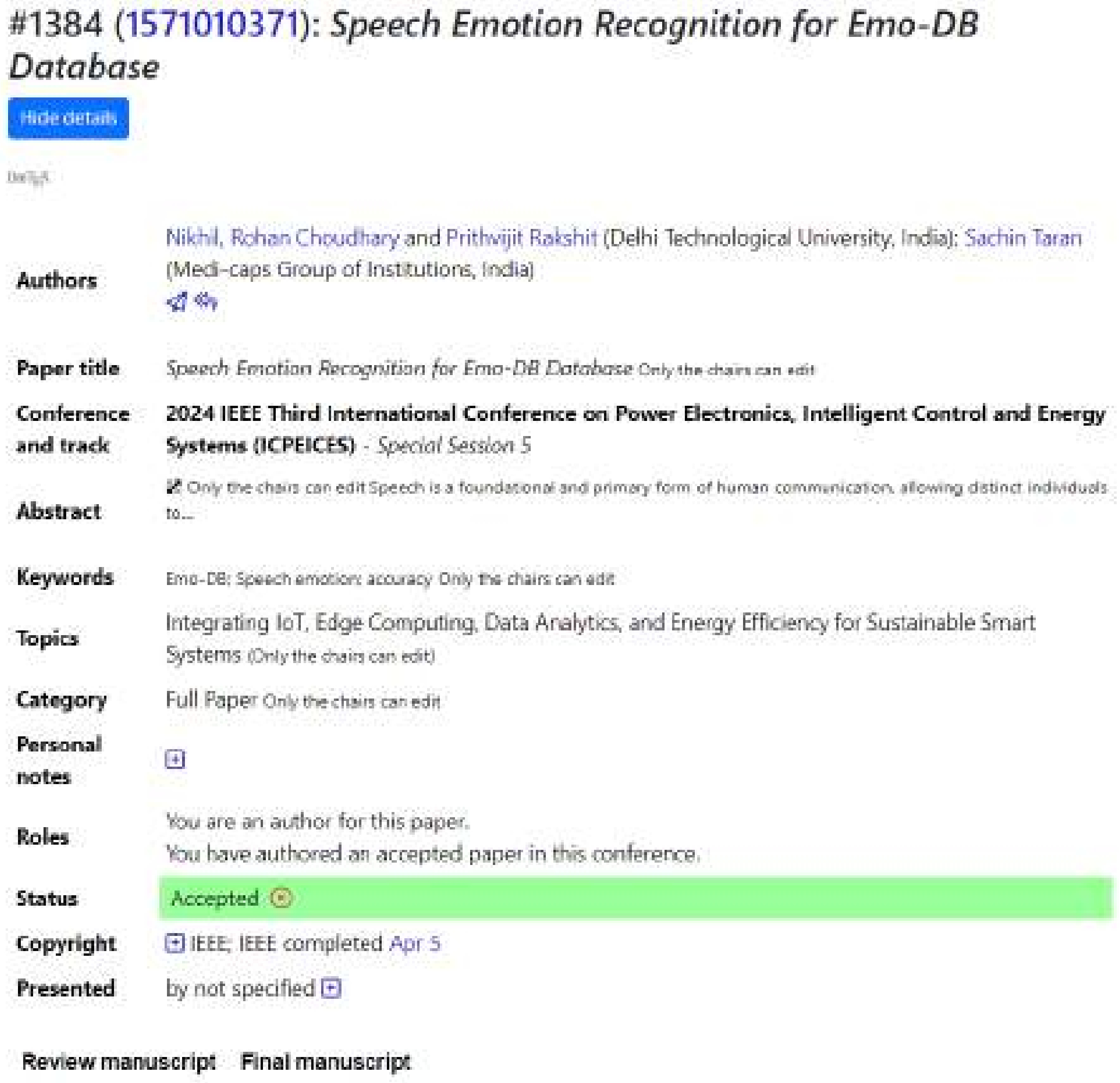
[30] T. Tuncer, S. Dogan, and U. R. Acharya, "Automated accurate speech emotion recognition system using twine shuffle pattern and iterative neighborhood component analysis techniques," *Knowledge-Based Systems*, vol. 211, p. 106547, 2021.

LIST OF PUBLICATIONS

Paper Name: SPEECH EMOTION RECOGNITION FOR EMO-DB DATABASE

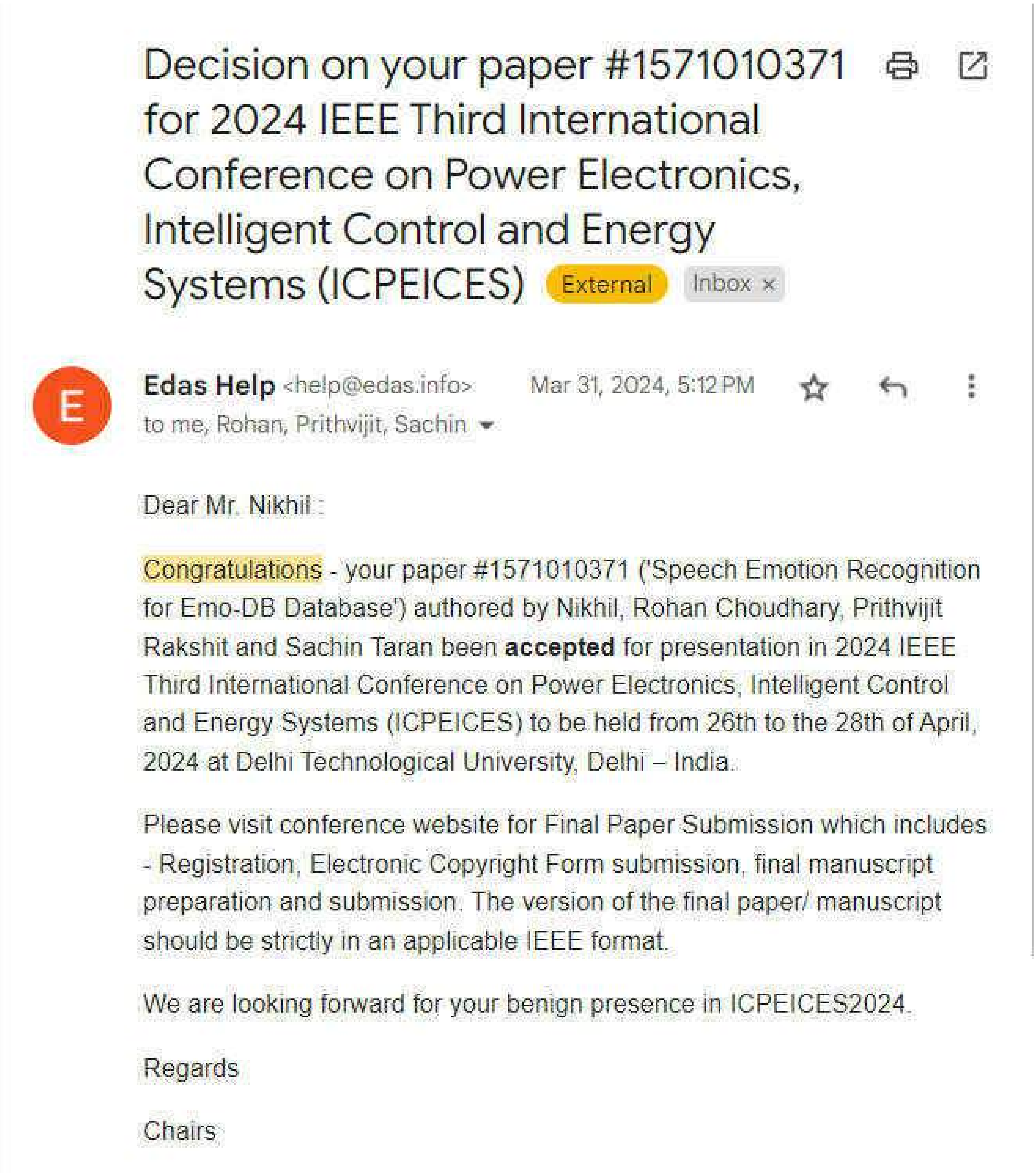
Publication Name: 3rd International Conference on Power Electronics, Intelligent Control, and Energy Systems (IEEE-ICPEICES-2024)

Status: ACCEPTED



#1384 (1571010371): *Speech Emotion Recognition for Emo-DB Database*

Authors	Nikhil, Rohan Choudhary and Prithvijit Rakshit (Delhi Technological University, India); Sachin Taran (Medi-caps Group of Institutions, India)
Paper title	*Speech Emotion Recognition for Emo-DB Database* Only the chairs can edit
Conference and track	**2024 IEEE Third International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES)** - *Special Session 5*
Abstract	Only the chairs can edit Speech is a foundational and primary form of human communication, allowing distinct individuals to...
Keywords	Emo-DB; Speech emotion; accuracy Only the chairs can edit
Topics	Integrating IoT, Edge Computing, Data Analytics, and Energy Efficiency for Sustainable Smart Systems (Only the chairs can edit)
Category	Full Paper Only the chairs can edit
Personal notes	
Roles	You are an author for this paper. You have authored an accepted paper in this conference.
Status	Accepted
Copyright	IEEE; IEEE completed Apr 5
Presented	by not specified



Decision on your paper #1571010371 for 2024 IEEE Third International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES)

Edas Help <help@edas.info> Mar 31, 2024, 5:12 PM
to me, Rohan, Prithvijit, Sachin

Dear Mr. Nikhil :

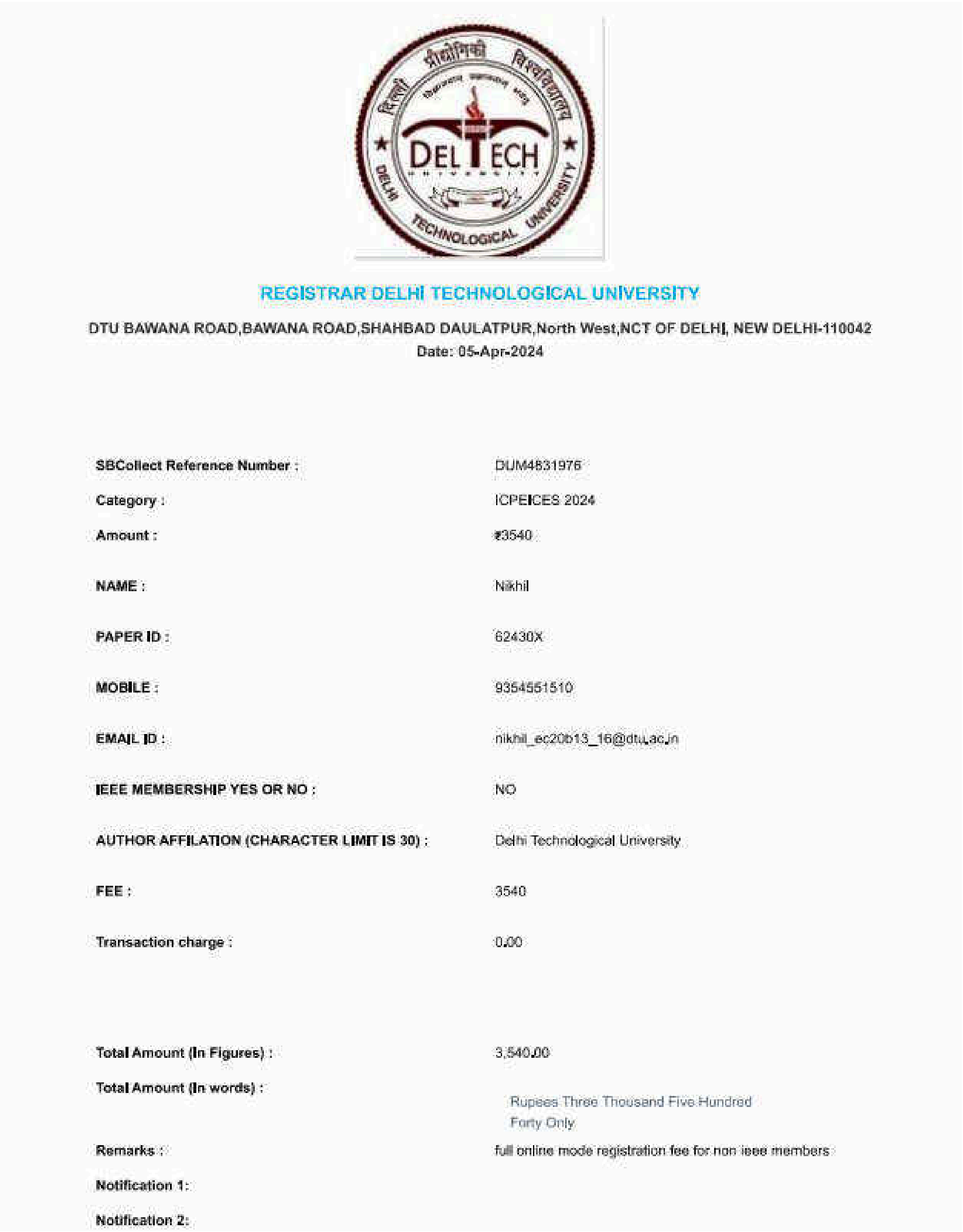
Congratulations - your paper #1571010371 ('Speech Emotion Recognition for Emo-DB Database') authored by Nikhil, Rohan Choudhary, Prithvijit Rakshit and Sachin Taran been **accepted** for presentation in 2024 IEEE Third International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES) to be held from 26th to the 28th of April, 2024 at Delhi Technological University, Delhi – India.

Please visit conference website for Final Paper Submission which includes - Registration, Electronic Copyright Form submission, final manuscript preparation and submission. The version of the final paper/ manuscript should be strictly in an applicable IEEE format.

We are looking forward for your benign presence in ICPEICES2024.

Regards

Chairs



REGISTRAR DELHI TECHNOLOGICAL UNIVERSITY

DTU BAWANA ROAD,BAWANA ROAD,SHAHBAD DAULATPUR,North West,NCT OF DELHI, NEW DELHI-110042
Date: 05-Apr-2024

SBCollect Reference Number :	DUM4831976
Category :	ICPEICES 2024
Amount :	₹3540
NAME :	Nikhil
PAPER ID :	62430X
MOBILE :	9354551510
EMAIL ID :	nikhil_ec20b13_16@dtu.ac.in
IEEE MEMBERSHIP YES OR NO :	NO
AUTHOR AFFILATION (CHARACTER LIMIT IS 30) :	Delhi Technological University
FEE :	3540
Transaction charge :	0.00
Total Amount (In Figures) :	3,540.00
Total Amount (In words) :	Rupees Three Thousand Five Hundred Forty Only
Remarks :	full online mode registration fee for non ieee members
Notification 1:	
Notification 2:	