Grenze International Journal of Engineering and Technology, June Issue

A Review on Speech Emotion Recognition with Machine Learning Techniques

Keerthana M[1], Gopalakrishnan K[2], Kavina C M[3], Kanishga S[4], Fharmaan A[5] and Kavitha K[6]
[1-5]Velalar College of Engineering and Technology / Department of IT, Erode, India
Email: k3931174@gmail.com, gopalakrishnanbtech@gmail.com, mohanasundaramkavina@gmail.com, kanishgasaravanan2@gmail.com, fharmaan02@gmail.com
[6]Kongu Engineering College / Department of CSE, Erode, India
Email: kavitha.kavi123@gmail.com

***Abstract*— The challenges of speech emotion recognition (SER) are examined in this study, highlighting limitations in conventional methods and the need for a deeper understanding. The review addresses the complexity of emotion definition and feature representation, crucial for accurate recognition. In essence, the research contributes to enhancing SER through a comprehensive exploration of machine learning techniques. A complicated area of affective computing, speech-to-text analysis (SER) requires a detailed examination of the emotions expressed in voice signals. Within the last decade, SER has become integral to Human-Computer Interaction and advanced speech processing systems. It identifies research gaps, particularly in fine-grained emotion classification essential for applications like psychological counseling. Addressing the uncertainty in emotion definition and the complexity of feature representation, the review paper advocates a deep dive into machine learning techniques to enhance SER's accuracy and effectiveness. This contribution aims to advance the understanding of speech emotion recognition, providing valuable insights for researchers and practitioners in this evolving field.**

***Index Terms*— Speech Emotion, Speech Emotion Recognition (SER), Speech, Machine Learning.**

I. INTRODUCTION

The discipline of Speech Emotion Recognition (SER) is concerned with the creation of technologies and algorithms for the purpose of recognizing and analyzes human emotions as they are communicated via speech. It is a subset of affective computing, a larger field that seeks to develop computers that are able to identify, comprehend, and react to human emotions. Applications like mental health monitoring, customer service, human-computer interaction, and the creation of more emotionally aware technology are where SER is especially useful. The primary objective of Speech Emotion Recognition is to identify and categorize the emotional states that a speaker conveys via their speech. Happiness, sorrow, rage, fear, surprise, and other emotions are examples of emotional states. Understanding the emotional content of speech can enhance the effectiveness of various applications, ranging from virtual assistants that respond empathetically to users, to systems that analyze emotional well-being based on speech patterns.

SER researchers and engineers use a range of methods to extract pertinent features from speech signals, including

Grenze ID: 01.GIJET.10.2.249_1

linguistic (semantic and syntactic information), prosodic (such as pitch, rhythm, and intensity), and spectral (related to the frequency content of the speech signal) features. Deep learning, a type of ML using deep neural networks, has particularly contributed to important advancements in tasks like image and speech recognition. Machine learning models, including deep learning approaches, are commonly used to analyze these features and classify the emotional states expressed in the speech. Despite significant progress, challenges remain in accurately recognizing and interpreting emotions from speech due to the variability in individual expression, cultural differences, and the complexity of emotional states. Ongoing research continues to refine and expand the capabilities of Speech Emotion Recognition systems, to create more robust and context-aware technologies that can better understand and respond to human emotions conveyed through speech.

II. Literature Survey

Predicting emotion in speech is neither simple nor easy. According to Li Min Zhang, Yu Beng Leau, Giap Weng Ng, and Hao Yan, [1] The difficulties in Speech Emotion Recognition (SER), a critical component of human-computer interaction with applications in senior care, healthcare, and education, are discussed in this work. Sufficient recognition accuracy is still a challenge, even with advances in feature extraction and model identification. In order to get around this, the authors provide a F-Emotion method for feature selection and a parallel deep-learning model for emotion recognition. To help with the best feature combination selection for recognition, the F-Emotion algorithm determines an F-Emotion value for every speech emotion feature. Two datasets (EMO-DB and RAVDESS) are the subject of the study, which yield impressive respective accuracy rates of 82.4% and 88.7%. Among the innovative contributions is the F-Emotion algorithm, which accurately evaluates the relationship between speech emotion aspects and emotion kinds. Using statistical techniques, it examines how each emotion type is distributed and clustered for every speech characteristic parameter. Based on the best feature combinations found by F-Emotion, a parallel deep learning model is developed that leads to a notable increase in recognition accuracy. It distinguishes between each emotion separately, in contrast to current models, and illustrates how various feature combinations affect various emotions. Additionally, a voting system for decision fusion is put into place, giving outputs from each parallel channel varying weights in order to get an overall result for emotion recognition. This all-encompassing method improves speech emotion recognition accuracy. By providing a methodical examination of the connection between emotion kinds and speech emotion characteristics, the work advances feature selection techniques in SER. Promising results are demonstrated by the suggested parallel deep learning model and F-Emotion algorithm, highlighting the significance of customized methods for differentiating between different emotions in voice inputs.

According to Syed Asif Ahmad Qadri, Teddy Suriya Gunawan , Taiba Majit Wani , Eliathamby Ambikairajah and Mira Kartiwi ,[2] Speech Emotion Recognition (SER) has proven essential to speech processing systems and human-computer interaction (HCI). Despite advancements, the quantitative and qualitative differences in how humans and machines interpret emotional aspects of speech pose challenges in merging knowledge from interdisciplinary fields. The paper synthesizes recent literature on SER design, highlighting the research gap and urging consideration from related researchers and institutions. Humans uniquely convey themselves through speech, a data-rich and culturally significant form of communication. While alternative methods like text and emojis exist, speech remains crucial. Emotion, gender, personality, goal, and state of mind that is frequently missed by conventional speech recognition algorithms are the paralinguistic information that speech contains. An ineffective understanding of paralinguistic features, especially in children, can lead to social skill deficiencies and psychopathological manifestations, emphasizing the need for coherent communication machines. Emotion recognition, historically focused on facial expressions, has expanded to include speech signals in recent times. SER aims to explore emotional states through speech signals but faces challenges in extracting effective emotional features. Classifying methodologies process and characterize speech signals to identify embedded emotions, requiring labeled data, preprocessing, and feature extraction. The choice of features significantly impacts classifier performance, with various models like Linear Discriminant Classifiers, Gaussian Mixture Models, and deep learning classifiers employed for emotion classification. The paper organizes its content to discuss SER systems, databases, and speech processing, feature extraction, and classification methods. Recent works, challenges faced by SER, and the paper's conclusions are outlined in subsequent sections. For emotion recognition ongoing research aims to enable programmable devices to engage in speech communication, focusing on designing robust methods. The analysis emphasizes the use of speech databases, prosodic and spectral acoustic features, and traditional and deep learning classifiers. Despite progress, challenges persist, requiring robust algorithms and efficient classification techniques for improved Human-Computer Interaction (HCI).

According to Shifang Cai, Ce Wang, and Gang Liu , [3] Speech signal processing's speech emotion recognition (SER) area is expanding, and data-driven research and advancements in processor capacity have made deep learning techniques more popular. However, challenges persist, such as limited datasets and weak emotion perception in existing networks. This research states that a novel method based on human-like implicit and emotional perception emotional characteristic categorization is proposed, drawing inspiration from brain science. The effectiveness of this method is validated by preliminary trials on the IEMOCAP dataset, which demonstrate a substantial increase in weighted accuracy (WA) by 3.17% and unweighted accuracy (UA) by 2.43%. SER involves machines automatically identifying human emotions from speech, crucial for human intelligence, decision-making, social interaction, and various applications like depression diagnosis and call centers. This paragraph emphasizes the vital role of emotion in human communication and outlines the practical applications of SER in diverse scenarios. The rise of deep learning has revolutionized speech emotion recognition, with neural network approaches dominating the field. From early use of recurrent neural networks (RNN) to recent applications of capsule neural networks and transformers, the performance has consistently improved. However, the major challenge: the network architectures used from natural language processing and computer vision. In addition to addressing the lack of datasets and emotion perception, the research emphasizes the need for improved emotional information modeling. The suggested method is depend on the structure of the human brain's emotion perception and is inspired by brain research. Using multitask learning, it presents an implicit emotion attribute categorization model akin to the human brain, exhibiting enhanced UA and WA on the IEMOCAP data set. The arrangement of the article is then described, the network design based on these features, the features of the perception of emotions by the human brain, the outcomes of the experiments, and the conclusions. In closing, it looks ahead to future developments in speech emotion detection, highlighting continuing research in brain science and the possibility of more progress in imitating cognitive emotion circuits.

According to Mark Teekit Tsun, Caslom Chua, Felicia Andayani, and Lau Bee Theng, [4] Speech-emotion recognition (SER) is very important for understanding human emotions in human-computer interaction. In this study, a unique hybrid model for SER is proposed that combines the advantages of Transformer and Long Short-Term Memory (LSTM) architectures. SER involves identifying emotions expressed through human speech, using machine learning (ML) architectures that rely on various features extracted from raw speech data. However, no single feature set has proven universally effective, leading researchers to combine multiple features for improved insight. The suggested hybrid model makes use of a Transformer encoder for multi-head attention on Mel Frequency Cepstral Coefficient (MFCC) feature vectors and LSTM to capture long-term relationships in speech signals. The hybrid model performs better, especially when it comes to identifying emotions in language-dependent datasets such as Emo-DB and RAVDESS. Recognition rates for RAVDESS and Emo-DB improved significantly compared to existing models, reaching 75.62% and 85.55%, respectively. The study suggests enhanced pre-processing methods, incorporation of additional feature types, and adaptation for real-time applications for further improvement. Data augmentation is proposed to address the challenge of data shortage in training datasets. Furthermore, the hybrid model's cross-corpus performance and its ability to recognize emotions in various languages are areas for future exploration.

According to Masato Akagi, Akira Sasou, and Bagus Tris Atmaja, [5] This research offers an evaluation of deep learning models with single-task learning and multitask learning methodologies for naturalness recognition and speech emotion. The emotion model uses five-point assessments for naturalness and arousal, valence, and dominance attributes—collectively referred to as dimensional emotion. Both naturalness scores (the auxiliary task) and dimensional emotion (the primary task) are concurrently predicted by multitasking learning. However, single-task learning predicts the naturalness score or the dimensional emotion separately. In terms of both naturalness predictions and dimensional emotion detection, multitask learning outcomes outperform the single-task learning research. In this study, single-task learning nonetheless performs better for naturalness recognition than multitask learning. In multitask learning, scatter plots of naturalness and emotion prediction scores against real labels show a model restriction, especially when it comes to predicting very high and very low scores. The MSP-IMPROV dataset helps naturalness, and a limited number of examples of unnatural speech is the reason for the inferior performance in naturalness prediction. The study highlights how future research on simultaneously predicting naturalness and emotion might enhance performance in emotion identification. The work makes a contribution by demonstrating how multitasking learning may manage naturalness scores and dimensional emotions at the same time, although with a little decrease in naturalness recognition performance. When compared to earlier research, the results are more accurate and reliable because to the 6-fold cross-validation examination. Future research directions include investigating acoustic features correlating with the naturalness of speech, addressing the gap between multitasking and single-task learning in naturalness recognition, and implementing a balancing strategy

to enhance model performance. Additionally, exploring continuous scores to ordinal label mapping and experimenting with alternative loss functions may further improve the overall results.
The authors discuss the broader implications of their work, emphasizing call-center analysis, spoken dialog systems, and education. Future directions include exploring the proposed method's effectiveness in different emotional speech corpora and real-world scenarios, showcasing the potential for widespread applicability.

III. Proposed System

The proposed system wants to increase the efficiency and accuracy of speech emotion recognition using convolutional neural network (CNN) with the vgg16 architecture. The vgg16 architecture, renowned for its effectiveness in image recognition tasks, is adapted to process speech spectrograms efficiently. Spectrograms capture the frequency content of speech signals over time and provide valuable insights into emotional states. The system pre-processes raw audio data to extract relevant features such as spectrograms. These features are then fed into the vgg16 CNN model for training and classification. Transfer learning techniques may be employed to leverage pre-trained weights of the vgg16 model, enhancing the training process and performance. The proposed system is evaluated using standard speech emotion datasets such as the Ravdess dataset. Several metrics for performance, like accuracy, recall, f1-score and precision are utilized to calculate the model's effectiveness in recognizing different emotional states.

A. Modules

Audio input data to spectrogram conversion : To convert audio input data into a spectrogram, the signal is first divided into small overlapping frames. Each frame undergoes a Fourier transform to obtain its frequency spectrum, which is then represented as intensity values in a two-dimensional grid, typically with frequency show in the vertical part of the axis and time on the horizontal part of the axis.

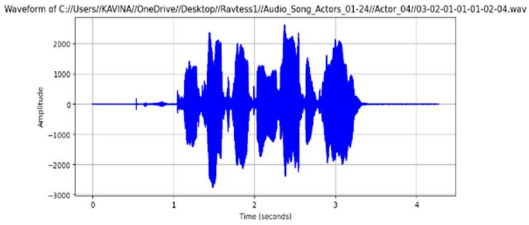

Fig 1 . Waveform of dataset (1)

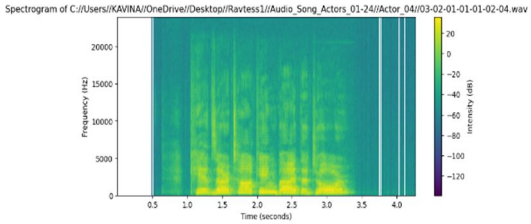

Fig 2 . Spectogram of dataset (1)

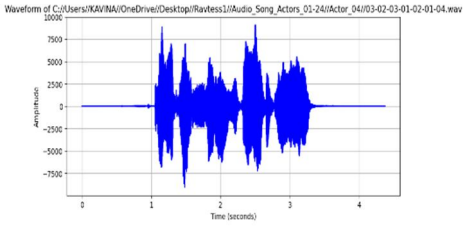

Fig 3 . Waveform of dataset (2)

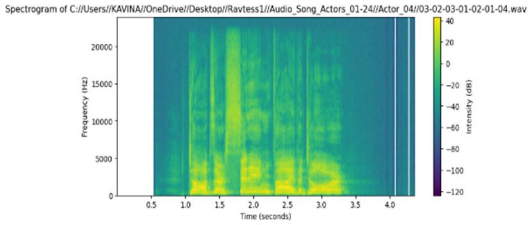

Fig 4 . Spectogram of dataset (2)

Model creation : The model summary provides insights into the architecture, including the number of parameters and the output shape at each layer, which is essential for understanding the model's complexity and behavior. It involves adapting the architecture and parameters of a pre-trained CNN, such as VGG16, to effectively process spectrogram inputs and classify emotions accurately.
Training & testing : Machine learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are trained on this exact data to know the patterns and correlations between the extracted features and the labeled emotions. During training, the model adjusts its parameters iteratively through back propagation to minimize a predefined loss function, thereby improving its ability to accurately classify emotions.
Performance metrics : Metrics quantify the model's ability to accurately classify emotions from speech signals. Commonly used metrics include precision, recall, F1 score, precision, and accuracy. Accuracy measures how reliable a model's predictions are, representing the proportion of properly classified samples in relevance to the total samples.

Model: "sequential_1"

Layer (type)	Output Shape	Param #
conv1d_3 (Conv1D)	(None, 40, 64)	384
activation_4 (Activation)	(None, 40, 64)	0
dropout_3 (Dropout)	(None, 40, 64)	0
max_pooling1d_2 (MaxPooling1D)	(None, 10, 64)	0
conv1d_4 (Conv1D)	(None, 10, 128)	41088
activation_5 (Activation)	(None, 10, 128)	0
dropout_4 (Dropout)	(None, 10, 128)	0
max_pooling1d_3 (MaxPooling1D)	(None, 2, 128)	0
conv1d_5 (Conv1D)	(None, 2, 256)	164096
activation_6 (Activation)	(None, 2, 256)	0
dropout_5 (Dropout)	(None, 2, 256)	0
flatten_1 (Flatten)	(None, 512)	0
dense_1 (Dense)	(None, 8)	4104
activation_7 (Activation)	(None, 8)	0

Total params: 209672 (819.03 KB)

Fig 5 . Model architecture summary (CNN-VGG16)

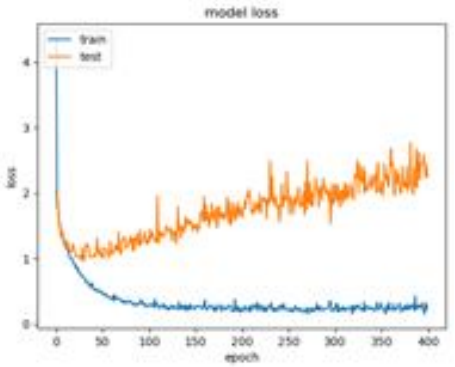

Fig 6 . Model loss

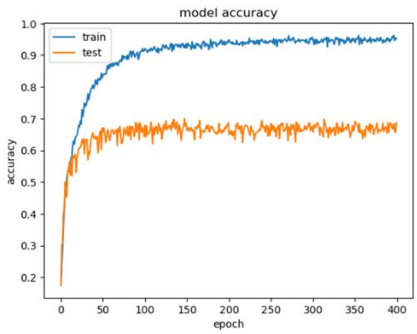

Fig 7 . Model accuracy

	precision	recall	f1-score	support
0	0.62	0.67	0.64	57
1	0.79	0.82	0.80	130
2	0.63	0.60	0.62	126
3	0.59	0.71	0.64	123
4	0.73	0.61	0.66	122
5	0.64	0.65	0.65	124
6	0.59	0.46	0.52	63
7	0.59	0.63	0.61	65
accuracy			0.66	810
macro avg	0.65	0.64	0.64	810
weighted avg	0.66	0.66	0.66	810

Fig 8. Performance Metrics

Expected Prediction : The confusion matrix serves as a valuable tool for diagnosing classification errors and refining speech emotion recognition models to achieve higher accuracy and robustness. The matrix is arranged in rows and columns. Every row denotes predicted class instances and every column represents actual class instances. Each cell in the matrix contains the count of instances that belong to the intersection of the predicted and actual classes.

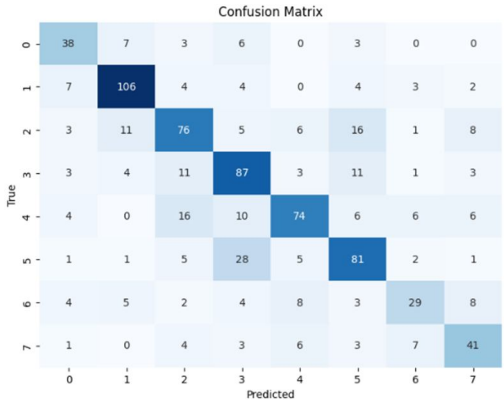

Fig 8. Performance Metrics

V. Conclusion

This review concludes by exploring the complexities of Speech Emotion Recognition (SER), highlighting the drawbacks of conventional techniques and underscoring the need for a deeper understanding. The research underscores the challenges in defining emotions and representing features accurately, crucial elements for precise recognition. By exploring various machine learning techniques, the survey strives to augment SER capabilities, recognizing its pivotal role in affective computing and its applications in Human-Computer Interaction and advanced speech processing systems over the past decade. Identifying gaps in research, especially in fine-grained emotion classification with potential applications in areas like psychological counseling, this work advocates for a nuanced approach. It tackles the uncertainty in emotion definition and the intricacy of feature representation through a comprehensive exploration of machine learning methodologies. The primary goal is to enhance the accuracy and efficacy of SER, providing valuable contributions to the evolving landscape of speech emotion recognition.

Acknowledgement

The authors wish to thank Department of Information Technology, Velalar College of Engineering and Technology, Erode for affording us the opportunity and encouragement.

References

[1] Li-Min Zhang, Giap Weng Ng, Yu-Beng Leau, Hao Yan, " A Parallel Model Speech Emotion Recognition Network Based on Feature Clustering" , DOI:10.1109/access.2023.3294274, pub month: July 2023.

[2] Taiba Majit Wani , Teddy Suriya Gunawan , Syed Asif Ahmad Qadri , Mira Kartiwi ,and Eliathamby Ambikairajah, "A Comprehensive Review of Speech Emotion Recognition Systems " doi: March 2021.

[3] Gang Liu, Shifang Cai, Ce Wang, "Speech emotion recognition based on emotion perception" access: EURASIP Journal on Audio, Speech, and Music Processing 2023, Article number: 22 (2023), Doi: 10.1186/s13636-023.

[4] Felicia Andayani, Lau Bee Theng, Mark Teekit Tsun, and Caslom Chua , "Hybrid LSTM-Transformer Model for Emotion Recognition From Speech Audio Files" access: https://doi.org/10.1109/ access. 2022.3163856, doi: March 2022.

[5] Bagus Tris Atmaja, Akira Sasou, and Masato Akagi, "Speech Emotion and Naturalness Recognitions With Multitask and Single-Task Learnings", DOI:10.1109/ACCESS.2022.3189481, Date of publication: January 2022.