IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, November 2024

A Comprehensive Review of Machine Learning Approaches for Speech Emotion Recognition

Trupti Dilip Kalokhe and Prof. Rashmi Kulkarni
Department of Computer Engineering
Siddhant College of Engineering, Pune, India
trupti.bhase18@gmail.com

Abstract: *Speech Emotion Recognition (SER) has become integral to enhancing human-computer interaction, leveraging advanced signal processing and machine learning techniques to analyze emotions in vocal expressions. This review highlights key methods such as Mel Frequency Cepstral Coefficients (MFCCs), Linear Predictive Cepstral Coefficients (LPCCs), and Perceptual Linear Predictive Coefficients (PLPCs) for feature extraction, alongside classification models like Support Vector Machines (SVM), Gaussian Mixture Models (GMM), and deep learning approaches such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). Recent advancements include hybrid models like Deep Belief Networks (DBN)-SVM and multimodal approaches combining speech, text, and facial features to improve accuracy. Applications of SER range from adaptive learning systems and mental health monitoring to real-time safety solutions. Despite progress, challenges such as noise resilience, limited dataset diversity, and emotion overlaps persist. Addressing these with strategies like transfer learning, autoencoders, and ensemble methods, the field continues to evolve toward creating scalable and reliable SER systems. Future research focuses on multimodal integration and refined architectures to enhance generalization and applicability in diverse scenarios.*

Keywords: Speech Emotion Recognition (SER), human-computer interaction, signal processing, machine learning, Mel Frequency Cepstral Coefficients (MFCCs), Linear Predictive Cepstral Coefficients (LPCCs), Perceptual Linear Predictive Coefficients (PLPCs), Support Vector Machines (SVM)

I. INTRODUCTION

Speech Emotion Recognition (SER) is a multidisciplinary research field that integrates elements of signal processing, machine learning, and human-computer interaction (HCI) to analyze and classify human emotions through speech signals. Emotions are a fundamental aspect of human communication, conveying intentions, context, and feelings that go beyond the spoken words. The ability of machines to recognize and interpret these emotions has become increasingly important with the rise of intelligent systems that aim to enhance user experience and interaction. SER has applications in diverse domains, such as mental health monitoring, adaptive learning systems, customer service, automotive safety, and entertainment, making it a critical area of innovation.

The significance of SER lies in its ability to process vocal signals, which naturally encode rich emotional information. Unlike other biometric modalities such as facial expressions or physiological signals, speech offers a non-invasive and highly accessible medium for emotion detection. Through variations in prosodic, spectral, and linguistic features, emotions can be effectively characterized and classified. By incorporating SER, systems such as virtual assistants, telehealth applications, and adaptive learning platforms can respond to human emotions with greater sensitivity and personalization. For instance, in the automotive industry, SER is being used to monitor driver states, detecting fatigue or anger to prevent accidents. Similarly, in healthcare, it aids in identifying emotional distress, facilitating early interventions.

The process of SER involves several interconnected stages, starting with preprocessing, where the speech signal is refined by removing noise, normalizing amplitude, and segmenting it into manageable units. This ensures the input quality necessary for accurate emotion detection. The next stage, feature extraction, involves identifying the most relevant attributes of the speech signal. Widely used features include Mel Frequency Cepstral Coefficients (MFCCs),

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, November 2024

which mimic the human ear's perception of sound, as well as Linear Predictive Cepstral Coefficients (LPCCs) and Perceptual Linear Predictive Coefficients (PLPCs). These features capture important acoustic and spectral properties of speech. Other prosodic features such as pitch, energy, and duration further add to the emotional characterization of the speech. Recent advancements in feature extraction have introduced methods like Bag-of-Audio-Words (BoAW) and dual-channel spectrograms, which enhance the representation of emotional nuances in speech signals.

Classification, the final stage, assigns emotional labels to the extracted features using machine learning algorithms. Traditional methods like Support Vector Machines (SVMs), Gaussian Mixture Models (GMMs), and Hidden Markov Models (HMMs) have formed the foundation of SER systems. However, deep learning models, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have significantly advanced the field by automating feature extraction and capturing complex patterns in the data. Hybrid approaches, such as Deep Belief Networks (DBN) combined with SVMs, have also shown promise in improving classification accuracy by leveraging the strengths of different algorithms. These advancements have made SER systems more robust and capable of handling the complexities of real-world applications.

Over the years, SER has witnessed significant evolution driven by advancements in computational power, algorithmic sophistication, and the availability of diverse datasets. Early systems relied heavily on handcrafted features and traditional classifiers, which limited their accuracy and adaptability. The advent of deep learning has revolutionized SER by enabling systems to learn directly from raw data, eliminating the need for manual feature engineering. Hybrid and ensemble models have further enhanced performance, combining traditional techniques with deep learning to achieve higher accuracy and robustness. Multimodal emotion recognition, which integrates speech with other modalities like text and facial expressions, represents the next frontier in SER, offering improved reliability and a more comprehensive understanding of human emotions.

Despite these advancements, SER still faces several challenges that hinder its widespread adoption. One of the primary challenges is the sensitivity of speech signals to noise, which can distort emotional cues and compromise system accuracy. Developing robust preprocessing techniques and noise-resistant models is crucial to overcoming this limitation. Another significant challenge is the lack of culturally and linguistically diverse datasets. Many widely used datasets, such as RAVDESS and EMO-DB, are limited in their representation of global languages and accents. Expanding datasets to include a broader range of emotional expressions and linguistic variations is essential for creating generalized systems.

The inherent complexity and overlap of human emotions pose additional challenges for SER. Traditional discrete emotion models often fail to capture the nuanced nature of emotions, leading to misclassifications. Continuous emotion models, which represent emotions in a multidimensional space based on parameters such as valence and arousal, offer a more flexible framework for emotion recognition. Furthermore, the interpretability of deep learning models remains a critical issue, particularly in sensitive domains like healthcare. Black-box models, while accurate, lack transparency, making it difficult to understand their decision-making process. Efforts to enhance model explainability are vital for building trust in SER systems.

The integration of SER into real-time systems presents another layer of complexity. Applications such as virtual assistants and automotive safety systems require algorithms capable of rapid inference without compromising accuracy. Balancing computational efficiency with performance remains an ongoing challenge. Addressing these limitations through innovative solutions such as transfer learning, data augmentation, and advanced feature extraction techniques is essential for the future growth of SER.

Looking ahead, the future of SER lies in the development of multimodal emotion recognition systems that combine speech with other biometric signals such as facial expressions, text, and physiological data. These systems promise to provide a holistic understanding of human emotions, enhancing the accuracy and reliability of emotion detection. Advances in deep learning architectures, such as attention mechanisms and transformers, are expected to further improve the ability of SER systems to focus on salient features of speech signals, even in complex and noisy environments. The integration of SER with emerging technologies like the Internet of Things (IoT) and edge computing will enable real-time applications in smart homes, healthcare, and transportation.

In conclusion, Speech Emotion Recognition represents a transformative technology with the potential to revolutionize human-computer interaction. By leveraging advanced signal processing techniques and machine learning models, SER

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, November 2024

systems have made significant strides in understanding and interpreting human emotions. While challenges such as noise sensitivity, dataset diversity, and emotion complexity persist, ongoing research and innovation are paving the way for more accurate, robust, and scalable systems. As the field continues to evolve, SER is poised to play a critical role in shaping the future of intelligent systems and emotional intelligence in technology.

II. LITERATURE SURVEY

Girija Deshmukh et al. in [1] proposed a system in which they obtained audio samples of Short-Term Energy (STE), Pitch, and MFCC coefficients in frustration, happiness, and sadness of emotions. Open source North American English served as expression and as feedback was used to record natural speech. Thus, only three emotions i.e., anger, happiness and sadness were recognized. They also identified the speaker's detailed features, such as sound, energy, pitch. The whole Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset is manually split into train and test sets. The multi-class Support vector machine (SVM) takes feature vectors as input, which is turned up as a model corresponding to each emotion.

Peng Shi in [2] introduced discrete model and continuous model of speech emotion recognition; different characteristics are analysed to make better description of emotions. When compared to Artificial Neural Networks (ANNs) and support vector machines (SVMs), the Deep Belief Networks (DBNs) have about 5% higher accuracy rate than the traditional methods. The output shows that the features which are extracted by Deep Belief Networks is much better than the original feature. DBN-SVM had slightly improved result than DBN-DNN because SVM classifies in small size better. DBN converts empty characteristics into deep abstract characteristics, resulting into better classification.

J. Umamaheswari et al. in [3] presented pre-processing being carried out using K-Nearest Neighbour (KNN) and Pattern Recognition Neural Network (PRNN) algorithms while the feature extraction was explained using a descended structure covering Gray Level Co-occurrence Matrix (GLCM) and Mel Frequency Cepstral Coefficient (MFCC). The outcomes were compared for their precision rate, accuracy and f-Measure with standard algorithms like Hidden Markov Model (HMM) and Gaussian mixture models (GMMs) were recognized as a better production than the benchmark algorithms. The emotional waves generate a pattern, the pattern of the signal is later recognized by PRNN. The believable nearest pattern concerning the signal is determined by K-NN approach. M.S.

Likitha et al. in [4] observed recognition requires assessment of the verbal communication wave to classify the required feeling, based on the training of its characteristics, like Sound, format, phoneme. On the side of withdrawal of functionality and examination, A good number of algorithms were made of a speech signal. The acoustic precision of the communication kinesics is a feature. Withdrawal of features is the process of removing a compact amount of information from the voice signal employed to reflect each speaker later on. Most Methods of extraction are at one's fingertips but the widely used method is coefficient (MFCC). Feature is the sound representative of the speech signal. An action that mines a little quantity of information from the verbal expression that can subsequently be used to act for each speaker is called feature extraction.

Zhang Lin et al. in [5] conveyed that SER innovation is used to monitor the driver's irregular emotions, and makes use of particular phrase recognition innovation to pick out the parking guidance in emergency situations. Three types of speech features are extracted that are prosodic, spectral-based and quality features. Frequently used characteristic parameters in speech recognition are the Mel-cepstral coefficient (MFCC) and the Linear Predictive Cepstral Model (LPCC). SVM is used for extracting the features. The concern of panic in the driver's voice can be defined by way of this system as an emergency situation. The parking instructions are listed on that basis. Since the voice parking guidance database and the speech emotion database used in this paper are collected on various settings, the recognition efficiency is significantly degraded as soon as the emotion of the parking guidance voice is declared.

Asaf Varol et al. in [6] expressed how sound is defined as a pressure wave that arises from the vibration of substance's present in a molecule. The investigation had gone through sound energy and its characteristics. More effective results are drawn from experiments using the speech signal spectrogram and Artificial Neural Networks (ANNs). Using EMO-DB dataset, SER uses role extraction techniques such as acoustic analysis and analysis of spectrogram to perform SER. In these current trends, they have also discussed the growing scope of SERs like in the field of signal processing, pattern recognition, psychiatry. Also, for yielding higher success rates, the author speaks different machine learning methods are to be performed on various kinds of datasets having various kinds of tests.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, November 2024

Abhijit Mohanta et al. in [7] have analysed emotions such as angry, fear, happy, neutral from emotional speech signal where features such as loudness, detecting voiced region, excitation energy were used for analysis. They call these features as sub-segmental features. Using features such as, instantaneous fundamental frequency (F0) using Zero Frequency Filtering (ZFF), signal energy, formant frequencies, and dominant frequencies the analysis of these feelings has been done. The study analyses generation characteristics of four different emotion states and not the classification of emotional states portrayed by the actor. For locating instantaneous F0 and zero-crossing rate (ZCR), some Signal processing methods, ZFF and STE were used, respectively.

Edward Jones et al. in [8] considered Speech emotion recognition as an exciting ingredient of Human Computer Interaction (HCI). The main approach for SER must be feature extraction and feature classification. Linear and non-linear classifiers can be used for Feature classification. In linear classifiers, frequently used classifiers are Support Vector Machines (SVMs), Bayesian Networks (BN). Since, Speech signal is considered variating, thus, these types of classifiers work effectively for SER. Deep learning techniques possess more advantages for SER when compared to traditional methods. The deep learning techniques does not require manual feature extraction and tuning, also, it has capability to detect complex structure with.

Michael Neumann et al. in [9] presented their conclusions illustration that gaining knowledge on unlabelled voice entities can be appropriate for Speech Emotion Recognition (SER). They have used t-distributed neighbour embeddings (t-SNE) to analyse visualizations of different representations. However, no divisible clusters are found in the 2D projections. These plots are excluded as of capacity they require. The autoencoder is trained on a large dataset. They have incorporated representations generated by autoencoders, which, in turn leads to steady developments in identification accuracy of SER model. The research also gives us a way to discover and experiment different alternatives of autoencoders and study procreative adversarial networks for representation learning.

Radim Burget et al. in [10] used German Corpus (Berlin Database of Emotional Speech) data which had more than 250 recordings. Each recording was distributed into 20 millisecond segments avoiding the overlapping. Then 3098 silent segments were eliminated to cut up the information into training, validation and testing sets. For removing the silent parts of audio Google WebRTC voice exercise detector was used in pre-processing. And then all the files are standardized so that they have zero mean and module variance. The input data given to Deep Neural Networks (DNN) were presented in batches with each of 21 iterations. Each batch contain a similar number of divisions and pattern is succeeded with pattern of different divisions like neutral, angry, sad and so on. DNN had no perception of the actual experience of what the actor is conveying, nor had any perception of the nature.

Kunal Bhapkar et al. in [11] provided a comprehensive survey of Speech Emotion Recognition (SER) techniques, highlighting the importance of preprocessing, feature extraction, and classification in SER systems. The paper categorized SER models into two types: discrete speech emotion models, which classify speech into distinct emotions, and continuous models, which represent emotions as points in a multidimensional space. The authors emphasized the significance of using Mel Frequency Cepstral Coefficients (MFCC) for feature extraction, which mimics the human ear's sensitivity to different frequencies. Classification methods discussed include traditional classifiers like SVM, GMM, and KNN, as well as deep learning-based models like Deep Belief Networks (DBN) and Deep Neural Networks (DNN). The survey also addressed challenges in SER, such as the selection of appropriate datasets, handling noise in speech signals, and improving classification accuracy through advanced feature extraction techniques. Despite its breadth, the paper lacked experimental results and detailed performance comparisons of the discussed methods.

Babak Basharirad and MohammadrezaMoradhaseli in [12] critically analyzed various Speech Emotion Recognition (SER) methods, focusing on feature extraction, classification techniques, and dataset challenges. They highlighted the importance of selecting robust features, such as Mel-Frequency Cepstral Coefficients (MFCCs), Linear Predictive Cepstral Coefficients (LPCCs), and Perceptual Linear Predictive Coefficients (PLPs), which are commonly used in SER. Recent advancements, including auditory-inspired long-term spectro-temporal features, were also discussed. Classification approaches like Support Vector Machine (SVM), Hidden Markov Models (HMM), Neural Networks (NN), Gaussian Mixture Models (GMM), and K-Nearest Neighbor (KNN) were reviewed. The authors emphasized the potential of hybrid methods, such as combining SVM with Radial Basis Function (RBF), to improve recognition accuracy. The study evaluated datasets like the Berlin Emotional Speech Database and AIBO database but criticized their limited cultural and linguistic diversity. Challenges in SER, such as overlapping emotions within a single speech

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, November 2024

signal and variability in language and accent, were identified. The paper concluded by advocating for the use of ensemble techniques and deep learning architectures for future SER advancements.

Prof. Kinjal S. Raja et al. in [13] utilized two datasets, Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and Toronto Emotional Speech Set (TESS), to analyze emotions like anger, fear, sadness, and happiness from speech. RAVDESS included 1440 samples covering eight emotions such as calm, happy, and sad, recorded by 24 actors, while TESS provided 2800 samples from two women portraying seven emotions. During preprocessing, speech samples were filtered to remove background noise and environmental variations, ensuring clearer datasets. Feature extraction methods like Mel Frequency Cepstral Coefficients (MFCC), Log Mel-Spectrogram, Chroma, Spectral Centroid, and Spectral Rolloff were used, with MFCC offering frequency mapping closer to human auditory perception. Feature selection was performed by extracting global features such as Min, Max, Mean, and Standard Deviation from MFCC data, reducing redundancy and enhancing model efficiency. Classification algorithms like SVM, Random Forest, k-NN, and Neural Networks (RNN and LSTM) were applied to predict emotions. The study showcased spectrograms for various emotions and emphasized SER's potential in applications such as adaptive learning systems, vehicle safety, and autism support. While promising, the work highlighted the need for more diverse datasets to improve system generalization.

SamanehMadanian et al. in [14] conducted a systematic review on Speech Emotion Recognition (SER) using machine learning techniques, highlighting three core steps: data preprocessing, feature extraction, and emotion classification. For preprocessing, they emphasized noise and silence removal using methods such as Google WebRTC Voice Activity Detector. Feature extraction focused on prosodic, spectral, and deep features, with MFCCs being widely used. Feature selection methods like PCA and Fisher Criterion were used to enhance model accuracy by removing irrelevant data. They reviewed classification models including SVM, RNN, and ensemble models like CRNN. Challenges such as speaker dependency and data imbalance were addressed using techniques like Cross-Speaker Histogram Equalization and data augmentation. The study advocated for the use of ensemble models and neural network-based algorithms for better performance, proposing a design combining deep feature extraction with advanced classifiers for robust SER systems.

In [15], This paper investigates the effectiveness of perceptual-based speech features for emotion detection, employing features such as MFCCs, PLPC, MFPLPC, BFCC, RPLP, and IMFCC. The algorithm utilizes deep neural networks (DNN) to evaluate auditory cues and identifies predominant features containing significant emotional information. The study validates the algorithm on the Berlin database, considering seven emotions in 1-dimensional categorical space and 2-dimensional continuous space with valence and arousal dimensions. Comparative analysis indicates substantial improvement in emotion recognition performance using DNN with the identified combination of perceptual features. The importance of emotion expression and perception in human interaction is emphasized, noting their influence on message interpretation and decisionmaking at the cognitive level. The paper highlights the need for man-machine interfaces to recognize user emotional states and react accordingly. The challenging task of emotion detection in the rapidly advancing field of humanmachine interaction (HCI) is acknowledged, with applications extending to household robots, interactive gaming, mobile communication, call centers, and more, integrating emotion recognition as an essential feature.

The [16] This paper delves into the challenges and recent advances in detecting emotions from vocal audio, primarily focusing on speech signals. Emotion recognition from speech remains a complex issue in artificial intelligence, necessitating the careful selection of a suitable classification model for effective detection. The document reviews various features and extraction techniques related to emotional speech data, providing a comprehensive analysis of different classification methods with a specific emphasis on classifying diverse emotions within the speech detection model. Significant features like MFCC, fundamental frequencies, and LPCC for emotion classification are discussed, underscoring their importance. The paper highlights potential applications of speech emotion recognition in teaching, entertainment, and clinical diagnosis, emphasizing its role in improving human-to-machine communication. As organizations increasingly demand effective communication interfaces between people and machines, the paper stresses the importance of automating emotion detection from speech for understanding consumers and making informed decisions. Despite significant progress, the review acknowledges the ongoing efforts needed to make machines more

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, November 2024

human-friendly and adept at accurately understanding emotions, particularly as human dependence on machines is expected to grow in the next 10-20 years.

This paper [17] tackles the intricate task of emotion detection from voice signals, crucial for improving human-computer interaction (HCI). It surveys existing literature on speech emotion recognition, emphasizing the use of traditional speech analysis methods and the recent emergence of deep learning strategies. The study explores various databases, emotions collected, and contributions to speech emotion recognition, with a specific focus on the research team's Speech Emotion Recognition Project. Employing Python 3.6, the RAVDESS dataset, and Pycharm, the project utilizes unsupervised learning, specifically the MLP-Classifier algorithm, to build a model adept at recognizing frustration or annoyance in speakers' voices. This work contributes valuable insights to the growing field of emotion recognition in voice signals, particularly regarding deep learning applications and their impact on HC.

This [18] paper underscores the significance of speech in human communication and emphasizes the crucial role of emotion recognition for effective humanmachine interaction (HMI). Recognizing emotions in speech enhances the naturalness of communication and is particularly important for simulating realistic interactions in the virtual world. The paper provides a review of existing methods for speech emotion detection, highlighting the predominant use of features such as MFCC and Energy in current approaches. It acknowledges the historical importance of speech in human communication and discusses the increasing relevance of HMI in both industry and academia. The complexity of human brain analysis of auditory input is noted, emphasizing the conversion of sounds into conceptual ideas for instructions, commands, information, and entertainment. The paper also outlines the three main steps in speech processing: preprocessing, feature extraction, and pattern recognition, with a focus on the informative role of vowels in speech signals. Overall, the review sets the stage for further research in the field of speech emotion detection, recognizing its pivotal role in advancing natural and effective human-machine interactions.

This [19] paper emphasizes the integral role of emotions in human life, acting as a crucial means of expression and communication about one's well-being. It introduces the concept of Speech Emotion Recognition (SER), a system designed to extract and predict a speaker's emotional tone from audio signals. The paper discusses the categorization of emotions into types such as Anger, Happiness, Sadness, and Neutral, and highlights the significance of finite resources for developing an effective SER system. Spectral and prosodic features, including Mel-frequency Cepstral Coefficients (MFCC) and pitch, loudness, and frequency, are identified as crucial elements for training machine learning models to recognize emotional states in speech signals. The study presents a Speech Emotion Recognition system that surpasses existing models in data, feature selection, and methodology, aiming for improved accuracy (averaging 78%) and fewer false positives in identifying speech-based emotions. Overall, the paper underscores the importance of emotions in human communication and introduces advanced methodologies to enhance the precision of Speech Emotion Recognition systems.

A voice-based real-time emotion detection technique using recurrent neural network empowered [20] This paper addresses the increasing importance of mining audio data from human conversations, particularly in the context of the Internet of Things (IoT) and voice-based multimedia applications. Recognizing the substantial big data generated by these applications, reflecting various aspects of human behavior, the paper focuses on extracting emotions from speech as a critical element for next-generation artificial intelligence. Introducing a novel approach using Bag-ofAudio-Words (BoAW) for feature embeddings in conversational audio data, coupled with a state-of-the-art emotion detection model based on Recurrent Neural Networks (RNN), the study reports significant improvements in accuracy. The research highlights the significance of real-time multimedia applications in understanding human behavior and emotions through digital footprints.

Emotion detection from text and speech: [21] This survey explores the evolving field of emotion recognition, focusing on text and speech analysis as crucial sources of emotional expression. With people expressing emotions through various mediums such as writing, speech, and social media posts, the paper underscores the importance of efficient emotion detection for diverse purposes, including human-computer interaction and human-robot interaction in the context of growing robotic research. The survey reviews existing research efforts, emotion models, datasets, and detection techniques, emphasizing the multidimensional nature of human emotions. It acknowledges the challenges in accurately detecting emotions from textual content and highlights the increasing relevance of automatic emotion

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, November 2024

detection systems across various socioeconomic domains, offering a comprehensive overview of achievements and potential extensions for future advancements in the field.

English speech emotion recognition method based on speech recognition [22] This paper addresses the limited focus on English speech emotion recognition in current research efforts in China, emphasizing the importance of understanding emotional content in English texts. It highlights the significance of affective computing in intelligent human-computer interaction technology and its impact on various fields, including auditory and vocal mechanisms, pattern recognition, and artificial intelligence. The paper introduces an English speech emotion recognition method based on speech recognition, acknowledging its effectiveness while recognizing the need for further improvement. The limitations include a focus on experimental research with English databases and the necessity to extend research to encompass multiple languages. Additionally, the paper suggests exploring multi-modal information fusion, considering various human expressions such as facial expressions, words, and brain waves, to enhance overall accuracy in emotion recognition.

Deep features-based speech emotion recognition for smart affective services [23] This paper delves into speech emotion recognition using a deep convolutional neural network (CNN) with rectangular kernels applied to spectrograms, representing an innovative approach. Unlike traditional CNNs with square kernels designed for 2D image data, spectrograms encode information differently, with time along the x-axis and frequency along the y-axis. The proposed method introduces rectangular kernels of varying shapes and sizes, along with max pooling in rectangular neighborhoods, to effectively extract discriminative features. The study evaluates its performance on Emo-DB and Korean speech databases, showcasing superior results compared to many state-of-the-art techniques. The experiments involve training CNN models on spectrogram images and exploring a transfer learning approach. The paper highlights the effectiveness of this method in recognizing emotions from speech, particularly with its modification of kernel sizes and pooling strategies for better adaptation to spectrogram data.

Study on emotion recognition and companion Chatbot using deep neural networ [24] This research focuses on the development of a communication system incorporating speech emotion recognition and semantic analysis, with applications in companion robots, technology products, and medical purposes. The proposed system employs a Convolutional Neural Network (CNN) based on GoogLeNet to recognize five emotions from speech, achieving a top accuracy of 79.81%. Additionally, semantic analysis is performed using Recurrent Neural Network (RNN) in the Seq2Seq framework. The system comprises client and server components, with the client implemented on the Android system for user interaction and the server on Ubuntu Linux with a web server for backend processing. Users can record voice on the cellphone app, and the system analyzes emotions and conducts semantic analysis to function as a responsive chatting machine. Key contributions include the integration of Chinese word vectors for improved dialogue tolerance, direct emotion classification through CNN without tokenization, and the utilization of an app-based approach for user-friendly interaction and cost-effectiveness. The system's capabilities extend to collecting users' emotional indices for analysis by back-end care organizations. The experiments involve training on a high-performance computing system, and the proposed system demonstrates efficacy in recognizing emotions and conducting semantic analysis on speech data.

"Emotion Detection Through Speech Analysis" [25] The use of speech analysis techniques for emotion detection is examined in the research article "Emotion Detection Through Speech Analysis". To extract acoustic features, recognise emotional states from spoken language, and analyse language are some of the techniques it investigates. Machine learning algorithms are also discussed. In a variety of domains, including customer service, human-computer interaction, and mental health monitoring, the study examines the importance of emotion detection. Important discoveries include the ability of deep learning models to reliably identify emotions from voice inputs, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). To increase emotion identification ability, the paper also emphasises the need of taking into account both linguistic content and auditory cues. More issues are covered, including bias reduction, interpretability of the model, and shortage of data. As a whole, the study advances.

Emotion Detection from Voice-Based Classified Frame-Energy Signal Using K-Means Clustering". [26] A technique for identifying emotions from speech signals is presented in the publication "Emotion Detection Through Speech Analysis" The method entails breaking up the speech signal into frames and categorising the individual frames

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, November 2024

according to their energy levels. These frames are then grouped into groups that correlate to various emotional states using the K-Means clustering technique. The method deduces the main emotion in the speech signal by examining the distribution of frames among clusters. The preparatory stages of feature extraction, the specifics of the K-Means clustering algorithm's implementation, and the assessment of the method's effectiveness using pertinent metrics like accuracy and precision are probably covered in the paper. The study may also compare its methods to those already in use for detecting emotions and explore the possible.

"Speech Emotion Recognition Using a Dual-Channel Complementary Spectrogram and the CNN-SSAE Neural Network." [27] A innovative method for identifying emotions from voice data is presented in the publication "Speech Emotion Recognition Using a Dual-Channel Complementary Spectrogram and the CNN-SSAE Neural Network." The suggested approach makes use of a dual-channel complementary spectrogram representation, which records the speech signal's phase and magnitude information and offers an all-encompassing view of the audio data. Furthermore, a Convolutional Neural Network with Stacked Sparse Autoencoder (CNN-SSAE) architecture is presented by the authors to facilitate the extraction of high-level characteristics from the spectrogram representations. Efficient emotion identification is made possible by the CNN-SSAE model's ability to automatically extract discriminative characteristics from the input data. The construction of the dual-channel complementary spectrogram, the CNN-SSAE model's architecture and training process, and the assessment of the suggested strategy on benchmark data are probably covered in the work.

"Early Threat Warning Via Speech and Emotion Recognition from Voice Calls" [28] The paper "Early Threat Warning Via Speech and Emotion Recognition from Voice Calls" describes a novel method for using speech and emotion recognition in voice conversations to identify possible threats or crises. The suggested approach examines a number of speech signal characteristics, such as emotional content and acoustic characteristics, in order to spot patterns linked to potentially dangerous circumstances. With the use of machine learning methods like ensemble classifiers and deep neural networks, the system is able to automatically identify voice calls as either normal or suspicious. The method of feature extraction from voice signals, the creation and training of machine learning models, and the assessment of the system's effectiveness with the use of actual voice call data are probably covered in the paper. Outcomes might contain criteria like precision.

"A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face" [29] The paper "A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face" offers a thorough summary of the most current developments in deep learning-based emotion detection spanning voice, text, and facial expressions, among other modalities. The survey addresses several facets of multimodal emotion identification, such as deep learning architectures, feature extraction methods, data gathering strategies, and assessment measures. It talks about how combining data from several modalities might increase the resilience and accuracy of emotion identification. Along with examining these issues, the study also addresses potential remedies and future goals for research on issues including data fusion, model interpretability, and cross-modal inconsistency. The survey also offers a comparative review of current methods, emphasising their advantages and disadvantages. All things considered, the article is an invaluable resource for scholars and learning and ensemble approaches enhance generalisation and dependability, especially in situations with a dearth of labelled data. Moreover, models must be updated and adjusted continuously to retain performance in dynamic contexts. Interpretability, bias reduction, and real-world application remain obstacles in spite of advancements. In order to allow useful applications in domains like mental health, future research should concentrate on creating more interpretable models, guaranteeing equity across demographic groups, and improving deployment techniques.

III. SUMMARY OF THE LITERATURE SURVEY

This section compiles and synthesizes key findings from recent research on haze removal techniques, with a focus on advancements and challenges relevant to our project. As haze significantly impacts visual clarity and image quality, affecting various applications, researchers have explored a diverse array of methodologies—from traditional techniques such as dark channel prior and color attenuation to cutting-edge deep learning models—to enhance the precision and efficiency of haze removal. This review provides a systematic evaluation of these approaches, highlighting their

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, November 2024

effectiveness, adaptability, and limitations, while offering a comprehensive perspective on the progress and current state of haze removal methods in alignment with our project's objectives

Table: Summary of the Literature survey

Sr. No.	YOP	Title and Name of Author	Main Findings	Methodology	Limitations	Application
1	2018	Girija Deshmukh et al.	Recognized anger, happiness, and sadness using RAVDESS dataset.	Used SVM with manually split training and testing datasets.	Limited to three emotions and a single dataset.	Emotion detection for virtual assistants.
2	2018	Peng Shi	DBNs achieved 5% higher accuracy than traditional methods.	Applied DBN-SVM for classification and deep feature extraction.	Continuous models not extensively explored.	Improved SER for adaptive learning platforms.
3	2019	J. Umamaheswari et al.	Hybrid PRNN and KNN showed improved accuracy in emotion recognition.	Used hybrid PRNN and KNN algorithms with GLCM and MFCC.	Limited benchmark comparisons for hybrid algorithms.	Enhanced SER in healthcare applications.
4	2019	Sri Raksha R. Gupta et al.	MFCC identified as the most effective feature extraction method.	Extracted MFCC features for emotion classification.	Focused solely on MFCC without alternative features.	Speech-based emotion detection for educational tools.
5	2019	Zhang Lin et al.	SER used for monitoring driver's irregular emotions in emergencies.	Integrated prosodic and spectral features with SVM.	Recognition efficiency degraded due to dataset variability.	Emergency systems for automotive safety.
6	2019	Asaf Varol et al.	Spectrogram analysis provided effective results in SER.	Used spectrograms and ANN for emotion classification.	Did not address challenges in real-time applications.	SER for psychiatry and signal processing.
7	2019	Abhijit Mohanta et al.	Analyzed sub-segmental features for detecting basic emotions.	Analyzed ZFF and STE for sub-segmental feature extraction.	Limited analysis of classification results.	Basic emotion detection in telecommunication.
8	2019	Edward Jones et al.	Highlighted advantages of deep learning over traditional classifiers.	Used deep learning techniques for automatic feature extraction.	Did not address model interpretability challenges.	HCI applications for improved user interaction.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, November 2024

9	2019	Michael Neumann et al.	Autoencoders improved SER accuracy through unsupervised learning.	Trained autoencoders for unsupervised representation learning.	Required large datasets for effective training.	SER advancements in unsupervised learning systems.
10	2017	Radim Burget et al.	DNNs were used for classifying emotions with standardized data.	Segmented audio and trained DNNs on emotion categories.	No exploration of actor's actual emotional states.	Emotion classification for intelligent assistants.
11	2020	Kunal Bhapkar et al.	Survey of SER models emphasizing discrete and continuous models.	Reviewed classification techniques like DBN and DNN.	Lacked experimental results and performance comparisons.	SER for customer behavior analysis.
12	2021	Babak Basharirad et al.	Emphasized importance of robust features for SER accuracy.	Analyzed MFCC, LPCC, and PLPC for feature extraction.	Limited dataset diversity and cultural representation.	Improved SER for call center systems.
13	2023	Prof. Kinjal S. Raja et al.	Preprocessed datasets improved SER performance in diverse scenarios.	Applied MFCC, Chroma, and Log Mel-Spectrogram features.	Required more diverse datasets for generalization.	Vehicle safety systems with SER integration.
14	2023	SamanehMadanian et al.	Proposed ensemble models for robust SER systems.	Focused on ensemble models combining deep feature extractors.	Speaker dependency and data imbalance not resolved.	Robust SER systems for adaptive learning tools.
15	2023	Hema C, Marquez FP	CNN-based SER achieved improved accuracy in emotion detection.	Used CNN with deep learning for emotion detection.	Focused only on achieving higher accuracy.	Real-time SER for multimedia applications.
16	2023	SamanehMadanian et al.	Proposed ensemble models for robust SER systems.	Applied PCA and deep feature extraction techniques.	Speaker dependency and data imbalance not resolved.	Adaptive learning platforms with SER integration.
17	2022	Chamishka S et al.	Improved real-time SER using Bag-of-Audio-Words.	Used RNN with feature embeddings from BoAW.	Limited to voice data without multimodal integration.	Real-time SER for IoT-based applications.
18	2022	Liu M	Addressed English speech emotion	Proposed a speech	Focused on English, lacking	Speech emotion recognition in smart

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, November 2024

			recognition gaps.	recognition-based SER framework.	multi-language diversity.	assistants.
19	2021	Badshah AM et al.	Proposed CNN with rectangular kernels for spectrogram data.	Used CNN with rectangular pooling on Emo-DB dataset.	Tested only on spectrogram data from specific datasets.	Emotion detection for multimedia applications.
20	2021	Lee MC et al.	Developed a companion chatbot using GoogLeNet for SER.	Integrated semantic analysis with RNN and GoogLeNet.	Focused on semantic analysis; limited emotion classes.	Companion robots and healthcare tools.
21	2020	Johnson A	Highlighted acoustic features for emotion detection.	Applied CNN and RNN for feature extraction and classification.	Ignored linguistic content for emotion interpretation.	Call center and mental health monitoring.
22	2020	Hossain N et al.	Clustered speech frames to detect dominant emotions.	Segmented frames using K-Means clustering.	Limited exploration of emotion overlap challenges.	Basic emotion detection in telecommunication.
23	2018	Li J et al.	Introduced dual-channel complementary spectrograms.	Trained CNN-SSAE on spectrogram data.	Focused on spectrogram features without time-domain analysis.	Emotion recognition for smart affective systems.
24	2018	Ishtiak I et al.	Proposed early threat warning via speech analysis.	Applied machine learning on acoustic features for SER.	Tested primarily on controlled datasets.	Threat detection and emergency response systems.
25	2017	Lian H et al.	Surveyed deep learning-based multimodal SER methods.	Reviewed data fusion methods for multimodal systems.	Limited exploration of real-world applications.	Human-robot interaction in industrial automation.
26	2017	Rastogi R et al.	Explored traditional and emerging SER techniques.	Analyzed traditional and deep learning approaches.	Ignored explainability of deep learning models.	SER for customer behavior analysis.
27	2017	Vaishnav S et al.	Reviewed MFCC-based SER approaches.	Focused on MFCC and energy features for emotion detection.	Required diverse datasets for better generalization.	Educational tools leveraging SER capabilities.
28	2017	Hema C et al.	Proposed CNN-based emotional	Used CNN for emotional	Focused on increasing	SER for improving multimedia

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, November 2024

			speech recognition.	speech signal classification.	accuracy without real-world deployment.	experiences.
29	2017	Tripathi A et al.	Reviewed advances in speech emotion classification.	Emphasized classification models for emotion detection.	Did not address challenges in noisy environments.	Advanced SER systems for HCI applications.

IV. DISCUSSION

Speech Emotion Recognition (SER) has emerged as a crucial domain in human-computer interaction, enabling systems to interpret and respond to human emotions effectively. The findings from this review demonstrate significant progress in methodologies, from traditional machine learning approaches to sophisticated deep learning-based models. Each method brings unique advantages while also presenting specific challenges, shaping the trajectory of SER research.

Traditional approaches, such as Support Vector Machines (SVM), Gaussian Mixture Models (GMM), and Hidden Markov Models (HMM), have been foundational in SER. These methods rely heavily on handcrafted features like Mel Frequency Cepstral Coefficients (MFCCs), Linear Predictive Cepstral Coefficients (LPCCs), and prosodic features such as pitch and energy. While effective in controlled environments, these approaches struggle with scalability and robustness in real-world scenarios due to their sensitivity to noise, dataset diversity, and overlapping emotional expressions.

Deep learning models, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and hybrid architectures such as Deep Belief Networks (DBNs), have significantly improved the field. By learning hierarchical representations from raw data, these models eliminate the need for manual feature extraction, achieving higher accuracy and adaptability. Advanced techniques, such as attention mechanisms and transformers, have further refined SER by focusing on critical speech segments and capturing complex temporal dynamics. However, deep learning approaches often face challenges like computational demands, interpretability issues, and the need for large, annotated datasets

One of the recurring challenges in SER is the limited availability of diverse datasets that capture cultural, linguistic, and contextual variations in emotional expressions. Existing datasets, such as RAVDESS and EMO-DB, are instrumental but fall short in representing the complexity of real-world emotions. Expanding datasets to include diverse languages, accents, and scenarios is essential for building generalized SER systems.

Another critical aspect is the balance between accuracy and efficiency. While traditional methods are computationally lightweight, they often lack the precision required for modern applications. Conversely, deep learning models provide high accuracy but at the cost of increased computational resources, making real-time deployment challenging. Hybrid approaches that integrate the strengths of both paradigms are emerging as promising solutions, offering a balance between performance and efficiency.

The interpretability of SER systems, particularly deep learning models, remains an open research question. As these systems are increasingly used in sensitive domains such as mental health and automotive safety, understanding their decision-making process is critical for building trust and ensuring accountability.

V. CONCLUSION AND FUTURE SCOPE

Speech Emotion Recognition (SER) has made remarkable strides, transforming human-computer interaction by enabling systems to perceive and respond to emotional cues in speech. This review highlights the evolution of methodologies from traditional machine learning techniques, such as Support Vector Machines (SVM) and Gaussian Mixture Models (GMM), to advanced deep learning models, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). These advancements have significantly improved the accuracy and robustness of SER systems, making them more adaptable to diverse scenarios. However, challenges such as limited dataset diversity, computational efficiency, and model interpretability persist. Traditional methods, while efficient, lack the precision

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, November 2024

required for complex real-world applications. On the other hand, deep learning models, despite their superior accuracy, often require extensive computational resources and large, annotated datasets.

The integration of SER into various domains, including education, healthcare, automotive safety, and customer service, has demonstrated its transformative potential. Applications such as adaptive learning platforms, mental health monitoring tools, and emotion-aware virtual assistants have highlighted the value of SER in enhancing user experiences. However, as systems become more integrated into daily life, ensuring their fairness, reliability, and ethical alignment will be critical. The growing need for culturally and linguistically diverse datasets underscores the importance of inclusivity in SER research. In summary, while SER has achieved significant milestones, continuous efforts are needed to overcome existing limitations and unlock its full potential.

The future of SER lies in addressing its current challenges and expanding its application horizons. Developing multimodal emotion recognition systems that integrate speech with text, facial expressions, and physiological data will provide a more comprehensive understanding of human emotions. These systems can enhance the reliability and accuracy of emotion detection, particularly in complex or ambiguous scenarios. Advancements in transfer learning and data augmentation can mitigate the scarcity of diverse datasets, enabling the development of models that generalize well across different languages, cultures, and contexts. Furthermore, lightweight deep learning architectures optimized for edge computing will enable real-time emotion recognition in resource-constrained environments, making SER more accessible and scalable.

Ethical considerations will also play a crucial role in shaping the future of SER. Ensuring data privacy, minimizing algorithmic bias, and making models interpretable will be key to fostering trust and acceptance among users. Applications in mental health, autonomous systems, and education will benefit from these advancements, as reliable SER systems can improve decision-making and enhance user experiences. Additionally, the integration of SER into IoT and smart environments will unlock new possibilities, such as emotion-aware smart homes and personalized healthcare systems. Collaborative efforts across disciplines, including machine learning, psychology, and linguistics, will drive the evolution of SER, ensuring its alignment with societal needs. As SER continues to evolve, it is poised to redefine human-computer interaction and contribute to a more emotionally intelligent future.

REFERENCES

[1] G. Deshmukh, A. Gaonkar, G. Golwalkar, and S. Kulkarni, "Speech based Emotion Recognition using Machine Learning," Institute of Electrical and Electronics Engineers, Mar. 2019.

[2] P. Shi, "Speech Emotion Recognition Based on Deep Belief Network," Institute of Electrical and Electronics Engineers, Mar. 2018.

[3] J. Umamaheswari and A. Akila, "An Enhanced Human Speech Emotion Recognition Using Hybrid of PRNN and KNN," Institute of Electrical and Electronics Engineers, Feb. 2019.

[4] S. R. Gupta, M. S. Likitha, A. U. Raju, and K. Hasitha, "Speech Based Human Emotion Recognition Using MFCC," Institute of Electrical and Electronics Engineers, Mar. 2017.

[5] T. Kexin, H. Yongming, Z. Guobao, and Z. Lin, "Research on Emergency Parking Instruction Recognition Based on Speech Recognition and Speech Emotion Recognition," Institute of Electrical and Electronics Engineers, Nov. 2019.

[6] Y. S. Ü. Sonmez and A. Varol, "New Trends in Speech Emotion Recognition," Institute of Electrical and Electronics Engineers, Jun. 2019.

[7] E. Ramdinmawii, A. Mohanta, and V. K. Mittal, "Emotion Recognition from Speech Signal," Institute of Electrical and Electronics Engineers, Nov. 2017.

[8] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," Institute of Electrical and Electronics Engineers, Aug. 2019.

[9] M. Neumann and N. T. Vu, "Improving Speech Emotion Recognition with Unsupervised Representation Learning on Unlabeled Speech," Institute of Electrical and Electronics Engineers, May 2019.

[10] P. Harár, R. Burget, and M. K. Dutta, "Speech Emotion Recognition with Deep Learning," Institute of Electrical and Electronics Engineers, Feb. 2017.

[11] K. Bhapkar, K. Patni, P. Wadekar, S. Pal, R. A. Khan, and M. Shinde, "Speech Emotion Recognition: A Survey," International Research Journal of Engineering and Technology (IRJET), vol. 8, no. 3, pp. 922–925, Mar. 2021.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, November 2024

[12] B. Basharirad and M. Moradhaseli, "Speech Emotion Recognition Methods: A Literature Review," AIP Conference Proceedings, vol. 1891, 020105, 2017. DOI: 10.1063/1.5005438.
[13] K. S. Raja and D. D. Sanghani, "Speech Emotion Recognition Using Machine Learning," Educational Administration: Theory and Practice, vol. 30, no. 6(s), pp. 118–124, 2024. DOI: 10.53555/kuey.v30i6(S).5333.
[14] L. S. Tripathi, S. Tripathi, and D. Gupta, "Enhanced Speech Emotion Detection Using Deep Neural Networks," International Journal of Speech Technology, vol. 22, pp. 497–510, Sep. 2019.
[15] A. Tripathi, U. Singh, G. Bansal, R. Gupta, and A. K. Singh, "A Review on Emotion Detection and Classification Using Speech," in Proceedings of the International Conference on Innovative Computing & Communications (ICICC), May 2020.
[16] R. Rastogi, T. Anand, S. K. Sharma, and S. Panwar, "Emotion Detection via Voice and Speech Recognition," International Journal of Cyber Behavior, Psychology and Learning (IJCBPL), vol. 13, no. 1, pp. 1–24, Jan. 2023.
[17] S. Vaishnav and S. Mitra, "Speech Emotion Recognition: A Review," International Research Journal of Engineering and Technology (IRJET), vol. 3, no. 4, pp. 313–316, Apr. 2016.
[18] C. Hema and F. P. Marquez, "Emotional Speech Recognition Using CNN and Deep Learning Techniques," Applied Acoustics, vol. 211, Aug. 2023, 109492.
[19] S. Chamishka, I. Madhavi, R. Nawaratne, D. Alahakoon, D. De Silva, N. Chilamkurti, and V. Nanayakkara, "A Voice-Based Real-Time Emotion Detection Technique Using Recurrent Neural Network Empowered Feature Modelling," Multimedia Tools and Applications, vol. 81, no. 24, pp. 35173–35194, Oct. 2022.
[20] K. Sailunaz, M. Dhaliwal, J. Rokne, and R. Alhajj, "Emotion Detection from Text and Speech: A Survey," Social Network Analysis and Mining, vol. 8, pp. 1–26, Dec. 2018.
[21] M. Liu, "English Speech Emotion Recognition Method Based on Speech Recognition," International Journal of Speech Technology, vol. 25, no. 2, pp. 391–398, Jun. 2022.
[22] A. M. Badshah, N. Rahim, N. Ullah, J. Ahmad, K. Muhammad, M. Y. Lee, S. W. Baik, "Deep Features-Based Speech Emotion Recognition for Smart Affective Services," Multimedia Tools and Applications, vol. 78, pp. 5571–5589, Mar. 2019.
[23] M. C. Lee, S. Y. Chiang, S. C. Yeh, and T. F. Wen, "Study on Emotion Recognition and Companion Chatbot Using Deep Neural Network," Multimedia Tools and Applications, vol. 79, pp. 19629–19657, Jul. 2020.
[24] A. Johnson, "Emotion Detection Through Speech Analysis," Doctoral dissertation, Dublin, National College of Ireland.
[25] Y. N. Hossain, R. Jahan, and T. T. Tunka, "Emotion Detection from Voice-Based Classified Frame-Energy Signal Using K-Means Clustering," International Journal of Software Engineering & Applications (IJSEA), vol. 9, no. 4, pp. 37–44, Jul. 2018.
[26] J. Li, X. Zhang, L. Huang, F. Li, S. Duan, and Y. Sun, "Speech Emotion Recognition Using a Dual-Channel Complementary Spectrogram and the CNN-SSAE Neural Network," Applied Sciences, vol. 12, no. 19, pp. 9518, Sep. 2022.
[27] I. Ishtiak, M. M. Rahman, and M. R. Usmani, "Early Threat Warning Via Speech and Emotion Recognition from Voice Calls," Doctoral dissertation, BRAC University, 2020.
[28] H. Lian, C. Lu, S. Li, Y. Zhao, C. Tang, and Y. Zong, "A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face," Entropy, vol. 25, no. 10, pp. 1440, Oct. 2023..