IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, May 2025
Impact Factor: 7.67

ASR -Based Speech Emotion Classifiers

Mohanasundram A[1], Moses Samsan M[2], Omkar S[3], Yasvanthraj P[4], Vigneshwaran S[5]
Computer Science and Engineering[1-5]
Mahendra Institute of Engineering and Technology, Salem, India

Abstract: *Emotion recognition from speech signals plays a crucial role in Human-Machine Interaction (HMI), particularly in the development of applications such as affective computing and interactive systems. This review seeks to provide an in-depth examination of current methodologies in speech emotion recognition (SER), with a focus on databases, feature extraction techniques, and classification models. It has been done in the past using low-level descriptors (LLDs) like Mel-Frequency Cepstral Coefficients (MFCCs), linear predictive coding (LPC), and pitch-based features in methods like Support Vector Machines (SVM), Random Forests (RF), and Gaussian Mixture Models (GMM). But the development of deep learning techniques has completely changed the field. Models like convolutional neural networks (CNNs) and long short-term memory (LSTM) networks have shown that they are better at capturing the complex temporal and spectral features of speech. This paper reviews prominent speech emotion datasets, exploring their linguistic diversity, annotation processes, and emotional labels. It also analyzes the efficacy of different speech features and classifiers in handling challenges such as data imbalance, limited data availability, and cross-lingual variations. The review highlights the need for future work to address real-time processing, context-sensitive emotion detection, and the integration of multi-modal data to enhance the performance of SER systems. By consolidating recent advancements and identifying areas for further research, this paper aims to provide a clearer path for optimizing feature extraction and classification techniques in the field of emotion recognition.*

Keywords: *Emotion recognition*

I. INTRODUCTION

Speech Emotion Recognition (SER), the automatic identification and interpretation of emotions conveyed through speech signals, is a rapidly expanding field. Understanding and interpreting human emotions is fundamental to human communication and is essential in domains such as affective computing, human-computer interaction, psychological research, and therapeutic applications. By developing techniques and algorithms that enable machines to recognize and respond to emotional cues in speech, researchers aimThe associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang to enhance the connection between humans and intelligent systems, thereby improving user experiences, providing personalized services, and facilitating more effective communication .Emotional communication is a complex and multifaceted phenomenon, involving the expression of various emotional states such as happiness, sadness, anger, fear, surprise, and contempt. Specific vocal cue patterns and auditory characteristics present in speech signals can help identify these emotions, according to Scherer 2003 Vocal. For example, a cheerful voice typically exhibits a higher pitch, greater energy, and a faster speech rate, whereas a sad voice generally has a lower pitch, less energy, and a slower speech rate. Accurate emotion recognition from speech requires the identification and analysis of these auditory signals .

Despite significant advancements, challenges persist in voice emotion recognition. Background noise, inter-speaker variability, cultural influences, and the subjectivity of emotional expression all make it difficult to accurately classify emotions . To deal with these problems, we need strong feature extraction methods, advanced signal processing algorithms, and the addition of contextual information to make SER systems more accurate and resilient .

Current limitations in affective computing, especially in interpreting complex human emotions, are notable. SER technologies face difficulties in accurately capturing and analyzing subtle emotional variations due to limitations in existing algorithms and models. The integration of multimodal information, such as combining audio with visual or

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, May 2025
Impact Factor: 7.67

physiological cues, offers a promising avenue to address these limitations. Multi-modal approaches can provide richer emotional context, improving the robustness and accuracy of SER systems . For example, EEG-based emotion recognition, when combined with audio, addresses challenges such as masked emotions and provides a more comprehensive analysis . Additionally, real-time applications and the development of lightweight models are crucial for enhancing SER performance in resource-constrained environments. Future research should be done to improve context-awareness and emotional reasoning. This will help make SER models that are easier to understand, more flexible, and better able to handle changing emotional situations.

Conduct a comprehensive evaluation of various SER databases, including acted, elicited, and natural audio data. Assess each database based on criteria such as data type, modality, size, emotional categories, and language to determine their suitability and limitations for SER research.

Perform an in-depth analysis of both acoustic and non-acoustic speech features used in SER. Examine their effectiveness, purpose, and reported accuracy in capturing emotional states, and provide a comparative overview of their contributions to SER performance.

Review and compare traditional machine learning and advanced deep learning classifiers employed in SER. Evaluate their performance based on feature sets, accuracy, computational efficiency, and adaptability to various SER tasks.Identify and articulate current gaps in SER research, particularly in areas such as contextual understanding, multi-modal integration, and data privacy. Propose actionable recommendations and future research

II. SPEECH EMOTION RECOGNITION BEYOND HUMAN–COMPUTER INTERACTION

Research in the area of speech emotion recognition (SER) focuses on identifying and comprehending emotions presented in speech. SER was first created for human-computer interaction,butit hasnowgrowninto manyotherfields,going beyond the scope of conventional applications. It has been used in healthcare for diagnostic and evaluation reasons relating to mental health. SER assists contact centres in analysing client interactions, allowing businesses to improve their offerings. SER aids market research by measuring customer attitudes and preferences. For individualised interventions, SER in education offers insights into pupils' emotional states. Voice assistants, chatbots, entertainment, and gaming experiences are all improved with SER. Exploration of SER's potential in domains other than human-computer interaction is also ongoing.

A. APPLICATIONS OF SPEECH EMOTION RECOGNITION

In practical use-cases, we want to illustrate the broad application possibilities of voice-based SER. Most of the time, they are driven by a user-centric approach, with the intention of improving service quality or even quality of life. In certain cases, the approach is more system-centric, with the main goal being to enhance the functionality of other pattern recognition software.

1) MARKET RESEARCH

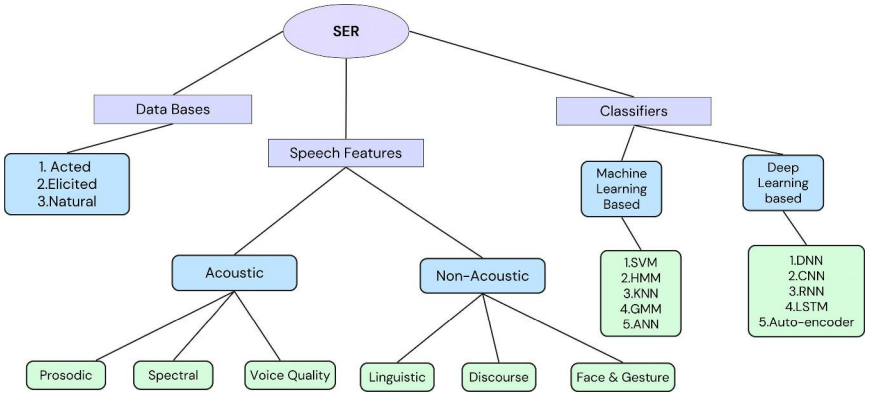
Business planning must include market research, and speech emotion recognition (SER) has become a useful tool in this area. A better understanding of consumer attitudes and emotional reactions to goods and services may be gained by analyzing customer feedback gleaned through surveys or recorded interviews. Researchers can identify underlying preferences, pain points, and reasons that influence consumer behavior by understanding the emotions portrayed in speech. Studies have demonstrated the use of SER in market research, showing how it may enhance product development, marketing initiatives, and client happiness [10], [11]. Its usein this area is still developing and offers fresh approaches to comprehending and interacting with target audiences.

2) MENTAL HEALTH MONITORING

A crucial component of healthcare is mental health monitoring, which evaluates and supports people's emotional well-being. Speech emotion recognition (SER) has become a potential method for keeping track of mental health thanks to technological improvements [7], [12]. For the benefit of mental health practitioners, SER may identify and analyze

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, May 2025
Impact Factor: 7.67

emotional states in real-time by examining speech patterns and extracting emotional elements. This allows for the quick identification of at-risk patients and offers them individualized care [13]. SER, like EEG-based emotion detection [7], also deals with the problem of accurately recording real emotional responses, which can be hard to do when someone has a condition like PTSD and their emotions may be hidden or changed [13]. The integration of SER in mental health monitoring holds the potential to transform mental healthcare, making it more accessible and boosting people's overall well-being [14].



3) LAW ENFORCEMENT AND SECURITY

The ability of speech emotion recognition (SER) technology to identify suspicious behavior and promote public safety has drawn significant interest from the fields of law enforcement and security [15], [16]. By analyzing speech patterns and emotional indicators in oral conversations, SER enables the detection of hostile, dishonest, or unusual behavior. SER algorithms [10], [17] can help law enforcement and security systems find threats faster, keep an eye on people more effectively, and lower the risks that might happen by using them [18].

In high-stakes environments such as airport security and border control, SER has demonstrated potential for realtime threat assessment. By integrating SER with other technologies like video analytics and face recognition, a comprehensive surveillance system can be developed that enhances threat identification and public safety [19]. However, practical implementations must address challenges such as the accuracy of emotion detection and the integration of SER into existing security protocols.

4) INTELLIGENT TUTORING SYSTEMS IN EDUCATION

Speech emotion recognition (SER) holds transformative potential in educational settings by providing insights into students' emotional states and enhancing learning experiences [20]. By measuring students' emotions in realtime, SER can offer educators critical information on engagement, motivation, and adjustment. This allows for the customization of teaching strategies, personalized interventions, and the creation of supportive learning environments [21], [22].

However, the use of SER in education raises important ethical considerations. Monitoring and evaluating student emotions necessitates careful attention to privacy and psychological impact [23]. The implementation of SER must include measures to protect students' privacy and ensure that emotional data is used responsibly. Additionally, potential psychological effects, such as increased stress or feelings of surveillance, should be assessed to avoid adverse outcomes. Ethical guidelines should be established to balance the benefits of SER with the rights and well-being of students [24].

5) VOICE ASSISTANTS AND CHATBOTS

As they offer personalized assistance, information, and entertainment, chatbots and voice assistants have firmly established themselves in daily life. These conversational agents have significant potential to enhance user engagement and experience through the integration of speech emotion recognition (SER) technologies [25]. SER enables chatbots

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, May 2025
Impact Factor: 7.67

and voice assistants to better perceive and respond to users' emotions, resulting in more engaging and empathetic interactions [26].

Sentiment analysis is a crucial application of SER in chatbots and voice assistants. These systems can adapt their responses and suggestions by instantly detecting and analyzing users' emotions. For instance, a chatbot might provide empathetic responses to resolve user dissatisfaction or offer relevant suggestions based on user excitement. This can improve user satisfaction and interaction quality [27], [28].

Additionally, SER can facilitatethecreationofemotionally sophisticated voice assistants and chatbots. By identifying emotions such as anger, sadness, or joy, these systems can adjust their tone, vocabulary, and behavior to better match users' emotional states, fostering more natural and humanlike interactions [29].

6) ENTERTAINMENT AND GAMING

Speech emotion recognition (SER) technology has been explored in the gaming and entertainment industries to enhance player experiences and provide more immersive entertainment [30], [31]. By analyzing players' speech patterns and emotional cues, game designers and entertainment companies can create dynamic and personalized experiences, adjusting gameplay, storylines, and interactions based on emotional states [32], [33].

Emotion-based gameplay is a significant application of SER in video games. SER can dynamically alter game elements such as difficulty levels and character behavior based on players' emotional reactions. For instance, games can offer hints or introduce new challenges if players show signs of frustration or boredom, or reward players displaying enthusiasm or joy. This enhances the overall gaming experience.

Furthermore, SER can improve the emotional impact of storytelling in entertainment mediums by enabling adaptive storytelling techniques. By analyzing viewers' or users' emotional responses, SER can facilitate dynamic scene modifications and branching storylines, leading to more immersive and engaging experiences [34], [35].

This review follows a systematic and organized approach to examining the advancements in Speech Emotion Recognition (SER), with a focus on databases, speech features, and classification techniques as explored in Sections III, IV, and V. The selection of reviewed studies and datasets was based on the categorization of audio data types—acted, elicited, and natural speech. The datasets were analyzed and compared across multiple parameters, including their type, modality (audio or multimodal), size, emotional categories, languages, and intended applications. This information was presented in tables to provide a clear overview of the different databases. Additionally, bar and pie charts were utilized to visually depict the distribution of reviewed studies across these datasets, highlighting trends in dataset usage for SER research.

In Section IV, the significance of different speech features for emotion detection was examined. Both acoustic features (such as pitch, energy, MFCCs, and spectral properties) and non-acoustic features (including prosodic and linguistic cues) were evaluated. The key studies, their employed features, intended purposes, and reported accuracies were summarized in tables to present a clear comparative analysis. Bar charts were also used to illustrate the frequency of feature usage across the studies, helping to highlight the effectiveness of various feature sets in the emotion recognition process.

Section V categorized the classification techniques into machine learning-based and deep learning-based methods. It was looked at how deep learning models, like convolutional neural networks (CNNs) and recurrent neural networks (RNNs), compare to traditional machine learning algorithms, like support vector machines (SVMs) and random forests. The tables provided a detailed comparison based on input features, goals, and performance metrics of these classifiers, while bar charts were used to visualize trends in classifier performance across different studies and datasets.

This review employs a systematic methodology to analyze Speech Emotion Recognition (SER) advancements, focusing on databases, speech features, and classification techniques. It uses strict inclusion criteria for peer-reviewed, relevant articles and excludes irrelevant studies. Data security was also considered, with an emphasis on informed consent and anonymization. The review features detailed tables and charts to categorize SER resources, revealtrends and gaps, and highlight the importance of data security, ensuring actionable and ethically sound findings for advancing emotion recognition technology.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, May 2025
Impact Factor: 7.67

III. DATABASES

In order to investigate and create algorithms to recognize and analyse emotions expressed via speech, databases are essential to speech emotion recognition research. With the use of these databases, which contain recordings of speech samples in a range of emotional states, researchers can develop and test machine learning models that are precise at identifying emotions. The use of these databases makes advancements possible in fields like affective computing, human-computer interaction, and psychological research.

In order to elicit a variety of emotional reactions, speech emotion recognition databases include recordings of human speech that have been meticulously crafted. To guarantee the veracity of the emotional content, these recordings are frequently performed by professional actors or people skilled in emotion induction techniques. These databases depict a wide range of emotions, such as joy, sorrow, anger, anxiety, and more. In order to investigate and create algorithms to recognize and analyze emotions expressed via speech, databases are essential to speech emotion recognition research. These databases include audio recordings of speech samples that display different emotional states, enabling researchers to train and test automated systems. The Berlin Database of Emotional Speech (EmoDB) and the EmotionalProsody Speech and Transcripts (EMO-DB) are notable databases utilized in speech emotion recognition research. While EMO-DB offers emotional voice recordings together with associated texts in German, EmoDB comprises of performed emotional speech recordings in the German language [36]. These databases provide thorough collections of emotional speech data, supporting research into the language and auditory characteristics connected to various emotional states. Another important source for speech emotion identification is the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [37]. It includes dyadic voice recordings of performers speaking in English while performing both scripted and spontaneous scenes. Researchers may investigate the dynamics of emotional interactions between people using the IEMOCAP database, which improves their understanding of emotional speech analysis.

The Surrey Audio-Visual Expressed Emotion (SAVEE) database [38]and the Toronto Emotional Speech Set (TESS) database [39] are important additional datasets. While SAVEE has emotional speech recordings by male actors in English, TESS features emotional speech recordings delivered by professional actors in English. These databases include various collections of emotional expressions, which helps researchers better understand how to recognize and perceive emotional speech. Databases for speech emotion recognition enable researchers to train machine learning models and algorithms, extracting acoustic, prosodic, and linguistic features of emotional states.

A. NATURE AND WAY OF COLLECTION OF RECORDINGS

Acted Speech Emotion database
Elicited Speech Emotion database
Natural Speech emotion database

1) ACTED SPEECH EMOTION DATABASE

In the study of emotion identification and analysis, acted spoken emotion databases are important tools. These databases are collections of voice recordings in which actors or other participants consciously mimic or act out particular emotional expressions in accordance with cues. These performed expressions make it possible to provide predictable and regulated emotional inputs for study and development [40].

The Berlin Database of Emotional Speech (EMO-DB) is a well-known illustration of a performed speech emotion database. Anger, fear, pleasure, boredom, melancholy, disgust, and neutral are only a few of the seven emotions that were asked of the professional actors that recorded German speech for this database. There are many distinct phrases and levels of intensity that fall under each category of emotion, offering a wide spectrum of emotional expressions. Both category and dimensional labels are included in the EMODB [36]. Databases of acted spoken emotions are used in study for a variety of reasons. They first offer a standardized and regulated data set for research and algorithm development on emotion identification. Researchers can guarantee consistent and reliable emotional inputs across several subjects and recordings by utilizing performed expressions. This makes it possible to compare and assess various methods for recognizing emotions more effectively [41].

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, May 2025
Impact Factor: 7.67

Second, acting databases make it possible to research the precise speech signals and acoustic characteristics linked to various emotional states. Prosody, pitch, intensity, and spectral features of the speech signal may be examined by researchers to find patterns and traits exclusive to each emotion category [42]. This study supports in the creation of more precise and reliable emotion identification systems and advances our knowledge of the auditory signals associated with emotional expression.

Additionally, databases of performed spoken emotions may be used to develop and evaluate machine learning models and algorithms. These databases support the creation of models that can automatically identify and categorize emotions in speech by supplying labeled data with wellknown emotional expressions.

It's crucial to remember that acted databases also have limits. These databases' phrases could fall short of accurately expressing the complexity and variety of emotions that people experience while speaking naturally and unprompted. Nevertheless, the development of efficient emotion identification algorithms and the advancement of our knowledge of emotional expression depend greatly on the use of performed spoken emotion databases [37].

2) ELICITED SPEECH EMOTION DATABASE

In the context of Speech Emotion Recognition (SER), an elicited database is one that contains speech samples that were gathered using deliberate, controlled elicitation approaches. Elicited databases are created expressly to elicit and record certain emotional states from the speakers, as opposed to spontaneous databases that record unforced emotional responses in real-world situations [43].

Participants in an elicited database are frequently told to display particular emotions or adhere to prepared situations that elicit the intended emotional responses [44]. Visual clues, textual directions, or spoken instructions can all be used to convey these instructions. Elicitation techniques are used to provide a regulated and uniform depiction of emotions in the database. Elicited databases are useful for SER research since they provide a number of benefits. First of all, they give researchers a controlled setting in which to examine and investigate certain emotional signals and traits like Egyptian Arabic speech emotion (EYASE) database is introduced that has been created from an award winning Egyptian TV series. The EYASE database includesfrom3maleand3femaleprofessionalactorsconsideringfour emotions: angry, happy, neutral and sad. Prosodic, spectral and wavelet features are computed from the EYASE database for emotion recognition [45]. Researchers can identify and study the auditory, prosodic, and linguisticfeaturesconnected to specific emotions by asking individuals to articulate them.

Second, elicited databases enable the systematic and repeatable collecting of a large number of speech samples demonstrating a wide spectrum of emotions. This makes it easier to design and assess SER models by allowing researchers to compile a varied dataset that encompasses a range of emotional states [46]

3) NATURAL SPEECH EMOTION DATABASE

Natural speech emotion databases record unplanned and impromptu emotional utterances in social settings. These databases,whichdifferfromacteddatabasesinthattheytryto record emotions as they emerge spontaneously during regular encounters, talks, or interviews. The emotional reactions in these databases come directly from the individuals and are not produced or acted out.

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database is an illustration of a natural speaking emotion database [37]. It includes recordings of dyadic encounters in which participants display a range of emotions while having scripted and natural dialogues. The database has audio and video recordings of speech, bodily motions, and facial expressions. Expert annotators who score the emotional content of the encounters provide the IEMOCAP database with the emotional labels. The MSPIMPROV database, which focuses on spontaneous emotional expressions in acting, is another noteworthy natural speech emotion database. The collection comprises of recordings of performers participating in improvised, unscripted scenarios in which emotions organically surface. The performers selfreport the feelings they felt during each recording, and these responses are used to create the emotional categories [47].

Natural speech emotion databases aid researchers in studying authentic emotional expression, analyzing social interactions, and context, enhancing emotion recognition systems' accuracy and robustness.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, May 2025
Impact Factor: 7.67

B. BASED ON MODALITIES

Audio-only Database

Multi-modal Databases

Text Based Databases

1) AUDIO-ONLY DATABASE

Speech emotion recognition audio-only databases isolate speech recordings from other modalities, such as visual or physiological information [48]. These databases offer an invaluable resource for researching the prosodic and auditory characteristics of emotional speech. Researchers can find patterns and discriminating signals that help people recognize and categorize emotional states by examining audio qualities including pitch, intensity, and spectral components.

2) MULTI-MODAL DATABASES

To give a complete representation of emotional expression, multi-modal databases for Speech Emotion Recognition (SER) combine many modalities, such as audio, visual, and physiological data. These databases give researchers the chance to investigate the merging of data from several modalities and create stronger models for emotion identification. An overview of multi-modal databases for SER is given below, accompanied with references:

Audio, visual, and textual modalities are all combined in the Multi-modal Emotion-lines data set. It comprises of movie speech that has been extracted together with utterancelevel emotional comments. The data set offers a comprehensive multi-modal resource for SER research by including audio signals, facial expressions, text transcriptions, and extra visual cues [49].

Another noteworthy multi-modal database is the Surrey Audio-Visual Expressed Emotion (SAVEE) database. It includes both audio and visual content, as well as spoken recordings that include male actors acting out seven various emotions. Researchers can investigate the fusion of auditory and visual signals for SER using the database's audio recordings and video clips of facial expressions [50].

Researchers have the chance to explore the complimentary nature of many modalities and create cutting-edge methods for emotion identification using these multi-modal datasets. Researchers may take use of a wider variety of characteristics and signals connected to emotional expression by merging auditory, visual, and other modalities, leading to more precise and reliable emotion identification models.

3) TEXT BASED DATABASES

Text-based databases are effective tools for managing and organizingtext-baseddata.Text-baseddatamaybeefficiently stored and retrieved using these databases' structured data storage.Relationaldatabasesanddocumentdatabasesaretwo popular categories of text-based databases [51], [52].

Tables with preset schemes are used in relational databases to hold data. Within these tables, text or var-char columns can be used to hold text data. Using main and foreign keys, relational databases link tables together. This makes it possible to efficiently query for and retrieve relevant textbased data.

On the other hand, document databases use documents that resemble JSON to store data in a flexible, schema-free way. These databases are perfect for storing textual material that is unstructured or partially organized. With the ability to execute full-text searches across documents, employ filters, and determine relevance scores, document databases provide extensive text search functionality.

References can be used in relational and document databases to define relationships between data elements. Document databases employ embedded documents or document references, whereas relational databases use foreign keys. The selection of a database is based on the particular requirements of the application, such as the type of data and anticipated query pattern [53]s.

Text-based databases provide strong mechanisms for textual information storage and retrieval, allowing for effective data organization and retrieval using specialized search engines like Apache Lucene and Elasticsearch [54], [55]

A survey of 26 studies on voice emotion identification highlights three major trends in dataset usage. Actor-based datasets were frequently utilized for training and evaluating models, providing structured insights into emotional dynamics. Natural datasets addressed real-world challenges such as speaker variability and environmental noise,

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, May 2025
Impact Factor: 7.67

focusing on unscripted emotions. Elicited datasets were used for controlled testing, bridging the gap between actor-based and natural data. Figure 2(a) and 2(b) illustrate the distribution of these datasets across different modalities: audio(Aud),audiovisual (AV), audio-video-text (AVT), and audio-text (AT).

In SER research, dataset availability significantly impacts usability and reproducibility. Open access datasets, such as AFEW and TESS, are freely available for academic purposes and encompass various data types, including actor-based, natural, and elicited data. Restricted access datasets, like MuSe-CAR and MSP-Podcast, require specific permissions and offer realistic emotion data captured under controlled conditions, which supports the study of genuine emotional expressions. Commercial datasets, such as those from the Linguistic Data Consortium, are available for purchase and provide extensive, high-quality data for in-depth analysis and model refinement. This variation in dataset accessibility influences the scope and depth of research in voice emotion recognition.

Database	Type	Modality	Size
CHEVAD (2017) [63]	Acted or Natural	Audio-Visual	140 min
Japanese Emotional DB (2021) [64]	Acted	Audio-Visual	100 min
Chinese Emotional DB (2021) [65]	Natural	Audio-visual	19,004 samples
Arabic Emotional DB (2021) [66]	Elicited	Audio	3280 samples
JL corpus (2018) [67]	Elicited	Audio	2400 recordings
MELD(2018) [68]	Acted	Audio-Video-Text	1400 dialogues and 14000 utterances

IV. CLASSIFIERS

Algorithms called Speech Emotion Recognition (SER) classifiers are made to automatically identify and categorise emotions in voice signals. The development of emotiona ware systems and applications, such as affective computing. human-computer interaction, and virtual agents, depends heavily on these classifiers. In SER, a variety of classifiers have been used, each with unique advantages and features. The capacity of Support Vector Machines (SVM) to identify an ideal hyperplane that divides several emotion classes in the feature space makes them well-liked. Hidden Markov Models (HMM) are a good choice for simulating the sequential character of emotions because they capture temporal

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, May 2025
Impact Factor: 7.67

relationships in speech signals. The probability distribution of feature vectors for each emotion class is represented using Gaussian Mixture Models (GMM).

Refrence	Feature	Purp
Polzeh(2011) [148]	Linguistic Features	The sify ditor datas cludi featu
Tzirakis(2017) [149]	Face and gesture features	Due that selve tion recog and netw
Haq(2009) [50]	Audio visual Features	In a this of cc and f
Zhao(2019) [150]	Facial and gesture features	For study tion globa tics f trogr netw ory built.
Pell(2009) [151]	Linguistic Features	In a tion lingu in th other

Due of their capacity to develop complex representations from unprocessed speech input, Artificial Neural Networks (ANN), especially deep learning models like Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), have been more popular in SER. An ensemble learning technique called Random Forests combines many decision trees to categorize emotions based on subsets of information. The amount of the dateset, available computing power, and required performance are all important considerations when selecting a classifier for SER. To determine which classifier is best for a certain SER job, researchers investigate and evaluate many classifiers. The goal of SER is to improve our comprehension and use of the emotional information that is communicated via speech by using these classifiers.

Classifiers are just as crucial to SER as speech characteristics. Now these may be divided into two groups:

(1) Classifiers based on Machine learning approaches. (2) Classifiers based on Deep learning approaches

Some researchers have employed hybrid methods, which combine standard and deep learning techniques. Although several classifiers have been tested for SER, it is still unclear which classifiers work the best. The literature of the classifiers is examined in this part using both conventional ML methodologies and cutting-edge DL approaches

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, May 2025
Impact Factor: 7.67

A. CLASSIFIERS BASED ON MACHINE LEARNING APPROACHES.

Due to its potential use in human-computer interaction, affective computing, and mental health evaluation, speech emotion recognition has drawn a lot of interest. Speech emotional classifiers powered by machine learning (ML) are effective methods for automatically identifying and classifying emotions expressed through speech. These classifiers identify significant characteristics that capture emotional cues using voice data and advanced algorithms. The feature classification process used in the ML techniqueallowsforthecollectionofspeechparametersincluding pitch, intensity, rhythm, and spectral content. SVMs, Random forests,HMMs,GMMs and ANNs are just a few of the classification techniques that may be used to train models that properly categorize speech into various emotional categories.

The requirement for automated and objective emotion identification systems is what spurs the development of speech emotional classifiers based on ML techniques. These systems may be used in a variety of fields, such as contact center analytics, virtual assistants, human-computer interface, and mental health evaluation. ML-based classifiers have the potential to improve human-machine interaction by precisely recognizing and deciphering emotional states from voice inputs. The paper will review the exiting work done using these classifiers:

1) SUPPORT VECTOR MACHINE(SVM)

SVMs are effective machine learning algorithms for voice emotion identification that also perform well on tasks involving classification and regression. They work especially well when dealing with multidimensional feature spaces and complicated decision boundaries. Using la-belled data linked with particular emotions, SVMs categorize voice samples into several emotional groups. They learn to maximise the margin between various classes by constructing an ideal hyperplane to divide various emotional classes in the feature space. SVMs are adept at extracting a variety of linguistic, prosodic, and acoustic properties from speech signals while handling high-dimensional feature spaces with ease.SVMsmayidentifynonlineardecisionboundariesusing this mapping, which captures intricate connections between features and emotions.

In 2005 Due to database architecture, sentiments, and various emotion sets, multiple emotion identification tests have been conducted [94]. This study uses three classifiers— spectral characteristics, prosodic features, and SVM—to examine automated emotion recognition in databases of Basque emotive speech. The first classifier, according to the results, has a 98.4% accuracy rate with 512 mixes, while the classifier using the top six prosodic characteristics has a 92.3% accuracy rate.

In 2011, Speaking utterances were categorized by the speech emotion detection system into one of five emotional states: disgust, boredom, sadness, neutral, or happy [97]. It makes use of samples from the Berlin emotional database as well as SVM, energy, pitch, LPCC, MFCC, and LPCMCC characteristics. The system's accuracy ranges from 66.02% with energy and pitch features to 70.7% with LPCMCC features to 82.5% with both.

and SVM outperforms [154] LDC, KNN, and RBFNN in identifying speech emotions, outperforming other techniques and achieving 85% accuracy in a corpus of emotional Chinese speech.

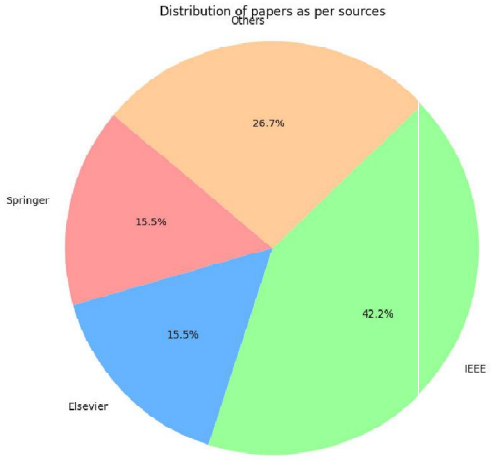
Low-level acoustic characteristics have demonstrated higher performance when combined with other features, such as lexical and statistical, or when using techniques like GMMs and SVMs. In 2017 it has been reviewed that a phoneme-based feature extractor called ConvLSTM-RNN is used in a voice emotion identification method [155]. It generates statistical features for use in SVM or LDA systems by outputting emotion probabilities to the input utterances. The strategy outperforms traditional ConvLSTMbased algorithms, according to experiments.

V. CRITICAL REVIEW AND OPEN PROBLEMS

The domain of Speech Emotion Recognition (SER) has experienced significant advancements, particularly through the application of deep learning techniques, such as Long Short-Term Memory (LSTM) networks, Convolutional Neural Networks (CNN), and hybrid models. These methods have yielded notable improvements, especially when

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, May 2025
Impact Factor: 7.67

S.no	Year	Reference	Features
01	2013	[191]	MFCC
02	2016	[192]	MFCC
03	2018	[193]	MFCC, LPC and other statis-tical features
04	2021	[194]	MFCC
05	2023	[195]	MFCC ,LFPC and other



(a) Distribution of papers as per sources

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, May 2025
Impact Factor: 7.67

Classifier	No.of papers	Reference	%
CNN	5	[196], [150], [197]. [198], [199]	17.86%
DNN	5	[200], [201], [202], [203], [204]	17.86%
RNN	6	[197], [100], [205], [206], [207], [208]	21.42%
LSTM	7	[209], [197], [210], [211], [212], [213], [214]	25%
Autoencoder	5	[215], [216], [217], [218]	17.86%

TABLE 22. Comparative performance metrics of traditional vs. deep learning approaches in SER.

Reference	Methods Compared	Accuracy	Speed	Computational Efficiency
Huang et al. (2014) [99]	DNN vs. ELM	DNN: Higher accuracy (5-10% better)	ELM: Faster than DNN	DNN: Higher computational cost
Mirsamadi et al. (2017) [11]	RNN with Attention vs. GMM	RNN: Higher accuracy (up to 10%)	GMM: Faster	RNN: Higher computational complexity
El Ayadi et al. (2011) [219]	SVM vs. DNN, CNN	DNN/CNN: Significantly higher accuracy	DNN/CNN: Slower due to depth	DNN/CNN: Higher resource demands
Zhao et al. (2018) [220]	Deep CNN vs. Traditional Methods	Deep CNNs: Superior accuracy	Traditional methods: Faster	Deep CNNs: Computationally intensive
Trigeorgis et al. (2016) [221]	Multimodal DNN vs. Traditional Models	Multimodal DNN: Significantly better accuracy	Traditional models: Faster	Multimodal DNN: Higher due to multimodal inputs
Poria et al. (2017) [222]	Various Deep Learning Approaches	Deep learning: Advancements in accuracy	Deep learning: Generally slower	Deep learning: Increased data and computational demands

The reliance on acted emotion datasets like RAVDESS, EMO-DB, and IEMOCAP has advanced emotion classification, yet these datasets often feature exaggerated expressions that do not capture real-world emotional subtleties, leading to reduced effectiveness in spontaneous speech. Additionally, there is a notable imbalance in the linguistic diversity of Speech Emotion Recognition (SER) datasets, with wellresourced languages such as English, Mandarin, and German being overrepresented, while low-resource languages like Kashmiri are underrepresented. This disparity limits the global applicability of SER models andunderscores the need for more inclusive datasets. Meanwhile, multimodal SER systems, which combine audio with visual and physiological data, show promise for

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, May 2025
Impact Factor: 7.67

improving emotion recognition in noisy environments. However, these systems are still developing and require further research to enhance the integration of different data streams and improve accuracy and robustness in real-world settings.

In summary, while advancements in deep learning architectures and traditional feature extraction methods have propelled the field forward, challenges surrounding generalizability, dataset diversity, and the integration of multimodal data remain pressing. Future efforts must focus on expanding the variety of speech datasets, embracing more sophisticated features, and optimizing multimodal approaches to ensure SER systems are more robust and applicable in real-world contexts. Following are some open problems:

Generalization to Real-World Scenarios: SER models need to be effective in naturalistic settings, requiring datasets that capture real-world emotional complexities and models capable of handling speech variability and noiseLow-Resourced Languages and Cultural Sensitivity: There is a need for SER systems that can recognize emotions in low-resource languages and across diverse cultural contexts. This includes creating annotated datasets and developing multilingual models.

Robustness to Noise and Speaker Variability: SER models must improve in handling noisy environments and speaker variability. Advancing noise-resistant feature extraction and robust pre-processing methods is crucial.

Incorporating Multimodal Data: Enhancing methods for integrating audio, visual, and physiological data is needed to achieve more accurate and nuanced emotion recognition.

REFERENCES

[1]. B. W. Schuller, *Intelligent Audio Analysis*, vol. 3. Berlin, Germany: Springer, 2013.

[2]. G. M. Dar and R. Delhibabu, ‘‘Exploring emotion detection in kashmiri audio reviews using the fusion model of CNN, LSTM, and RNN: Genderspecific speech patterns and performance analysis,’’ *Int. J. Inf. Technol.*, Aug. 2024, doi: 10.1007/s41870-024-02105-4.

[3]. C. Min Lee and S. S. Narayanan, ‘‘Toward detecting emotions in spoken dialogs,’’ *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 293–303, Mar. 2005.

[4]. J. Deng and F. Ren, ‘‘A survey of textual emotion recognition and its challenges,’’ *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 49–67, Jan. 2023.

[5]. S. Alisamir and F. Ringeval, ‘‘On the evolution of speech representations for affective computing: A brief history and critical overview,’’ *IEEE Signal Process. Mag.*, vol. 38, no. 6, pp. 12–21, Nov. 2021.

[6]. J. Zhang, Z. Yin, P. Chen, and S. Nichele, ‘‘Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review,’’ *Inf. Fusion*, vol. 59, pp. 103–126, Jul. 2020.

[7]. B. Chakravarthi, S.-C. Ng, M. R. Ezilarasan, and M.-F. Leung, ‘‘EEGbased emotion recognition using hybrid CNN and LSTM classification,’’ *Frontiers Comput. Neurosci.*, vol. 16, Oct. 2022, Art. no. 1019776.

[8]. Y. Dong, K. Zhao, L. Zheng, H. Yang, Q. Liu, and Y. Pei, ‘‘Refinement cosupervision network for real-time semantic segmentation,’’ *IET Comput. Vis.*, vol. 17, no. 6, pp. 652–662, Sep. 2023

[9]. A Esposito, A. M. Esposito, and C. Vogel, ‘‘Needs and challenges in human computer interaction for processing social emotional information,’’ *Pattern Recognit. Lett.*, vol. 66, pp. 41–51, Nov. 2015.

[10]. M. S. Hossain and G. Muhammad, ‘‘Emotion recognition using secure edge and cloud computing,’’ *Inf. Sci.*, vol. 504, pp. 589–601, Dec. 2019.

[11]. S. Mirsamadi, E. Barsoum, and C. Zhang, ‘‘Automatic speech emotion recognition using recurrent neural networks with local attention,’’ in*Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2227–2231.

[12]. Y. Yang, C. Fairbairn, and J. F. Cohn, ‘‘Detecting depression severity from vocal prosody,’’ *IEEE Trans. Affect. Comput.*, vol. 4, no. 2, pp. 142–150, Apr. 2013.

[13]. N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, ‘‘A review of depression and suicide risk assessment using speech analysis,’’ *Speech Commun.*, vol. 71, pp. 10–49, Jul. 2015.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, May 2025
Impact Factor: 7.67

[14]. F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller, ''Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data,'' *Pattern Recognit. Lett.*, vol. 66, pp. 22–30, Nov. 2015.

[15]. T. Bänziger and K. R. Scherer, ''Using actor portrayals to systematically study multimodal emotion expression: The gemep corpus,'' in *Proc. 2nd Int. Conf. (ACII)*, Lisbon, Portugal. Berlin, Germany: Springer, 2007, pp. 476–487.

[16]. S. Kaur and N. Kulkarni, ''Emotion recognition—A review,'' *Int. J. Appl. Eng. Res.*, vol. 16, no. 2, pp. 103–110, 2021.

[17]. S. T. Saste and S. M. Jagdale, ''Emotion recognition from speech using MFCC and DWT for security system,'' in *Proc. Int. Conf. Electron., Commun. Aerosp. Technol. (ICECA)*, vol. 1, Apr. 2017, pp. 701–704.

[18]. J. Tao and T. Tan, ''Affective computing: A review,'' in *Proc. 1st Int. Conf. (ACII)*, Beijing, China. Berlin, Germany: Springer, Oct. 2005, pp. 981–995.

[19]. C. M. Hurley, A. E. Anker, M. G. Frank, D. Matsumoto, and H. C. Hwang, ''Background factors predicting accuracy and improvement in micro expression recognition,'' *Motivat. Emotion*, vol. 38, no. 5, pp. 700–714, Oct. 2014.

[20]. S. Petrovica and H. K. Ekenel, ''Emotion recognition for intelligent tutoring,'' in *Proc. BIR Workshops*, vol. 1, 2016, pp. 1–9.

[21]. N. Banda and P. Robinson, ''Multimodal affect recognition in intelligent tutoring systems,'' in *Affective Computing and Intelligent Interaction*. Berlin, Germany: Springer, 2011, pp. 200–207.

[22]. F. Albu, D. Hagiescu, L. Vladutu, and M.-A. Puica, ''Neural network approaches for children's emotion recognition in intelligent learning applications,'' in *Proc. EDULEARN*, 2015, pp. 3229–3239.