FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 10 No. 3, February, 2026, pp 202 – 219
DOI: https://doi.org/10.33003/fjs-2026-1003-4565

STACKED LONG SHORT-TERM MEMORY AND HIDDEN MARKOV MODEL FOR SPEECH EMOTION RECOGNITION-A SYSTEMATIC REVIEW

[*1]Joshua Bature Hassan, [1]Temitayo Matthew Fagbola, [2]Lawrence Bunmi Adewole, [2]Daniel Dauda Wisdom

[1]Department of Computer Sciences Federal University Oye Ekiti, Ekiti State, Nigeria.
[2]Department of Computer Sciences, Federal University of Agriculture Abeokuta (FUNAAB), Ogun State, Nigeria.

*Corresponding authors' email: joshua.hassan@fuoye.edu.ng
ORCID ID: https://orcid.org/00-09000819334656

ABSTRACT

Speech Emotion Recognition (SER) plays a vital role in various real-world applications, from mental health diagnostics to human-computer interaction. This research examines the integration of Stacked Long Short-Term Memory (LSTM) networks and Hidden Markov Models (HMM) for SER, addressing the limitations of traditional models like Gaussian Mixture Models (GMM) and Support Vector Machines (SVM). While GMMs and SVMs offer decent performance, their inability to model temporal dynamics limits their accuracy in detecting complex emotions. The proposed hybrid model leverages LSTM's strength in handling sequential data and HMM's ability to model emotional transitions, making it highly suitable for real-world noisy environments. Preprocessing techniques such as MFCC and LPCC are applied to enhance feature extraction, and benchmark datasets like IEMOCAP and RAVDESS are used for evaluation. Finally, The Systematic review highlights the superior role of the hybrid model's performance on SER and sets the stage for a significant shift in future research in addressing bias and fairness in SER systems by combining LSTM and HMM as hybrid model. In recent studies, context-aware emotion recognition models are commonly evaluated using existing datasets such as IEMOCAP, MELD, DailyDialog, and SEMAINE, which provide conversational and multimodal emotional data. Features used often include acoustic features as MFCCs, spectrograms, prosodic features, linguistic features as BERT or transformer embeddings, and visual cues such as facial landmarks or video frames. Many models also incorporate contextual embeddings from transformers, graph attention networks, or sequential memory modules to capture speaker history and emotional shifts. Other findings showed that multimodal fusion significantly improves robustness, while training on noisy or real-life datasets with an improved generalization to real-world applications.

Keywords: SER, LSTM, HMM, GMM, Hybrid model

INTRODUCTION

Speech Emotion Recognition (SER) is an important and multifaceted field of research. It focuses on identifying and interpreting human emotions from speech signals (Coto-Jiménez, 2021). This capability extends beyond simple speech-to-text conversions or voice command systems, tapping into the rich emotional context embedded in spoken language. Emotions play a crucial role in human communication; they affect how we convey our thoughts, how others interpret our intentions, and how we form connections with others. Understanding emotions from speech can thus enrich human-computer interaction (HCI) and improve Artificial intelligence (AI) systems responsiveness as well as adaptability (Amami, 2023).

SER systems try to automatically recognize emotions expressed in a speaker's voice by analyzing the acoustic features and patterns in their speech. Unlike text-based emotion recognition which relies on analyzing the words used, SER focuses on non-verbal aspects of communication, such as tone, pitch, energy, and rhythm. These aspects often convey more emotional content than the actual words spoken (Shahin et al., 2023)

The general process of Speech Emotion Recognition (SER) involves three main steps namely signal preprocessing, feature extraction, and classification (Patel et al., 2024).

The raw speech signal is first cleaned of noise and distortions to ensure the data is in a usable form. This step is essential because poor-quality input can affect the accuracy of the emotion recognition process. Feature Extraction is the most crucial part of SER, where important patterns like pitch, loudness, and voice quality are extracted from the cleaned speech data. These patterns serve as indicators of the speaker's emotional state. After feature extraction, a classification model is used to predict the emotional state. Various models such as neural networks, Hidden Markov Models (HMM), and more recently, hybrid systems like Stacked Long Short-Term Memory (LSTM) combined with HMM are used for this purpose.

LSTM, a type of Recurrent Neural Network (RNN), is particularly effective because it can capture temporal dependencies in speech, making it ideal for emotion recognition tasks (Abbaschian et al., 2021).

Over the past decade, deep learning models have significantly enhanced the performance of SER, especially with LSTMs excelling in capturing sequential data like speech. These advancements have opened up SER to a wide range of applications across multiple fields (Ullah et al., 2023).

In Human-Computer Interaction (HCI), SER enables systems to guise a user's emotional state and adapt accordingly. For instance, virtual assistants like Siri, Google Assistant, and Alexa could improve interactions by recognizing emotions like frustration, joy, or confusion. This emotional awareness makes responses more empathetic, enhancing the overall user experience (Haque et al., 2023).

Call Centers and Customer Service, SER is used to detect emotions in real-time during customer interactions. If a customer is frustrated, the system can alert a representative to adjust their communication strategy. This improves customer satisfaction and reduces business churn rates. Additionally, SER systems can provide insights into areas where customer service teams need more support or training. SER has great potential for monitoring mental health conditions like

depression, anxiety, and stress. For instance, a system could analyze a patient's speech during therapy to monitor their emotional well-being. This early detection can allow for timely intervention, offering an additional layer of mental health care (Zhang et al., 2021).

Automotive Industry, SER in vehicles can monitor drivers' emotional states, detecting stress or fatigue. If a system senses that a driver is stressed, it could suggest calming music or offer navigation assistance, thereby improving road safety and reducing accident risks.

Entertainment and Gaming, in video games, SER can enhance player experiences by making in-game characters respond to the player's emotions in real-time. This creates a more immersive and interactive gaming environment, adapting the gameplay to the emotional state of the player (Alhasan et al., 2020).

Education, in online learning, SER can be used in intelligent tutoring systems to detect when students are confused or frustrated. The system could then adjust its teaching style or pace to better suit the student's emotional state, improving learning outcomes (Abbaschian et al., 2021).

SER is transforming how Artificial Intelligence (AI) interacts with humans by making systems more emotionally intelligent. This emotional awareness allows AI to make interactions more human-like, empathetic, and effective. From mental health monitoring to enhancing customer service, SER's ability to detect and respond to emotions in real time opens new possibilities for AI applications in healthcare, education, entertainment, and more (Huang et al., 2021).

as deep learning models such as Stacked LSTM and HMM continue to improve, the future of SER is bright, promising more accurate, capable systems that will integrate emotional intelligence into AI, leading to more natural and meaningful human-machine interactions. Fig.1. depicts, the 2D Emotional Space - based on Russell.

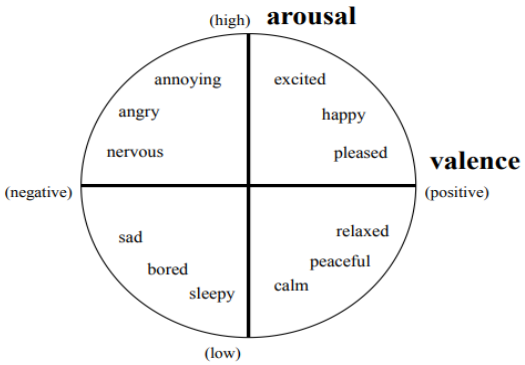

Figure.1: The 2D Emotional Space - based on Russell (Zhang et al., 2021)

SER are frameworks established to examine, classify, and represent human emotions according to psychological as well as physiological reactions (Bakker et al., 2014). They help scholars and machine learning (ML) models understand complex affective states in an organized manner, mapping emotions according to some precise dimensions or classes (Alhasan et al., 2020). Over time, various models such as discrete, dimensional, and componential have been proposed to identify emotional experiences. Among them, the Russell Circumflex Model of Affect stands out, representing emotions within a two-dimensional space defined by valence such as pleasant unpleasant; arousal such as activation or deactivation, allowing for a continuous interpretation of emotional states (Bakker et al., 2014).

From the Russell Model which shows when interacting with stimuli: affect, behavior, and cognition. In turn, these experiences lead to feeling, thinking, and/or acting - the three functions of the soul by Plato widely referred to in environmental psychology. Thus, based on a better understanding of the three dimensions of emotions and their relationship to other theories in the field, one can justify the replacement of the conventional two-dimensional model with the PAD space, including the third dominance axis - see Fig 2. (Zhang et al., 2021).

Emotion Models

Several emotion models are commonly used by researchers and authors to understand and represent human emotions:

i. Ekman's Basic Emotion Model Proposes six universal emotions (happiness, sadness, anger, fear, disgust, surprise) based on facial expressions and biological universality.
ii. Plutchik's Wheel of Emotions Organizes eight primary emotions (joy, trust, fear, surprise, sadness, disgust, anger, anticipation) in pairs of opposites, arranged in a circular structure to show emotion intensity and relationships.
iii. Russell's Circumplex Model of Affect Represents emotions in a two-dimensional space defined by valence (pleasant–unpleasant) and arousal (high–low activation).
iv. James-Lange Theory Suggests emotions result from physiological responses to stimuli; for instance, we feel afraid because we tremble.
v. Cannon-Bard Theory Argues that emotional experience and physiological response occur simultaneously rather than sequentially.
vi. Schachter-Singer (Two-Factor) Theory States that emotion arises from physiological arousal and cognitive interpretation of that arousal.
vii. Appraisal Theory (Lazarus) Emphasizes cognitive evaluation of events as the key determinant of emotional experience.
viii. These models collectively form the foundation for emotion recognition and analysis in psychology, neuroscience, and affective computing.

Challenges in Emotion Recognition From Speech

Emotion recognition from speech is a complex task, primarily because human emotions are conveyed subtly and differently. One of the keychallenges in Speech Emotion Recognition (SER) is variability in speech patterns. People express emotions differently based on their personality, age, gender, and even their situation. For instance, happiness might be

expressed with excitement in one person but with a calm tone in another.

A context-aware emotion detection system leverages on the external cues such as environmental information, speaker identity, or time-of-day in addition to primary signals such as speech to improve the accuracy of emotion classification (Lin et al., 2024). For example, CognEmoSense integrates speech data with contextual embedding (e.g., behavioral or situational metadata) and uses a transformer plus continual-learning module to adapt dynamically to individuals, achieving high accuracy (~88%) and robustness over time (Gendron & Guibon, 2024).

In speech-based systems, joint modeling of automatic speech recognition (ASR), speaker diarisation, and emotion recognition in a unified architecture ensures that the system understands not just "what is said" but "who says it" and "how they say it," significantly improving emotion detection in conversational contexts (Zhang et al., 2024)(Wang et al., 2025).

Moreso, multimodal transformer-augmented fusion methods combine speech and text features and utilize intra- and inter-modal attention to capture richer emotional representations and contextual dependencies, raising recognition performance respectively. By being sensitive to temporal, social, and situational context, such systems can more precisely reflect human emotional states (Zhao et al., 2025).

This makes it difficult for systems to generalize and accurately detect emotions across different speakers (Venkataramanan & Rajamohan, 2019).

Another major challenge is background noise. In real-world applications, speech often occurs in noisy environments—think of someone speaking on the phone in a busy street or during a meeting with multiple speakers. Distinguishing between the actual speech and background sounds like traffic or other conversations can reduce the accuracy of emotion detection models (Mohamed & Schuller, 2020).

One way to overcome background noise is to apply noise-reduction and speech enhancement techniques, such as spectral subtraction, Wiener filtering, or deep-learning-based denoising models that clean the audio before analysis (Chen & Zhang, 2024). Another approach is to use robust acoustic features and models, like MFCCs with noise compensation, RASTA filtering, or CNN/Transformer architectures trained on noisy datasets so the system learns to generalize. Additionally, multi-modal fusion (e.g., combining speech with facial expressions, physiological signals, or contextual cues) helps the system detect emotions correctly even when the audio quality is degraded (Chen & Zhang, 2024)(Nam & Park, 2024).

Cultural differences also play a significant role in how emotions are expressed. What might sound like anger in one culture could simply be a more passionate way of speaking in another. These factors create significant hurdles for designing universally effective SER systems. As SER moves toward real-world applications, addressing these challenges becomes essential to improve both accuracy and reliability (Tzinis & Potamianos, 2017).

Traditional Approaches to Speech Emotion Recognition

Before the advent of deep learning, Gaussian Mixture Models (GMM) and Support Vector Machines (SVM) were the go-to techniques for Speech Emotion Recognition. These models rely heavily on statistical methods to classify emotions based on extracted features from speech, such as pitch, loudness, and energy (Tanoko & Zahra, 2022)(Lin et al., 2024)(Chen & Zhang, 2024). GMM models worked by estimating the probability distribution of these features for different emotions, making decisions based on the likelihood of a certain emotion being present. While GMMs were effective to some degree, they struggled with more complex and variable emotional expressions in real-world data (Ahmed et al., 2023).

SVMs, on the other hand, focus on finding the optimal hyperplane that separates different emotion classes based on speech features. SVMs offered better performance than GMMs in many cases because they could handle non-linear data using kernel functions. However, both GMM and SVM approaches were limited by their dependence on handcrafted features and their inability to model temporal dependencies in speech. These models could analyze static snapshots of speech features but failed to capture how emotions evolve, which is critical for tasks like emotion recognition where context and timing matter. Despite these limitations, GMM and SVM laid the foundation for more advanced models that have since revolutionized SER (Jalal et al., 2019). Fig. 2. Depicts the Dimensional model of PAD as tripartite view of experience.

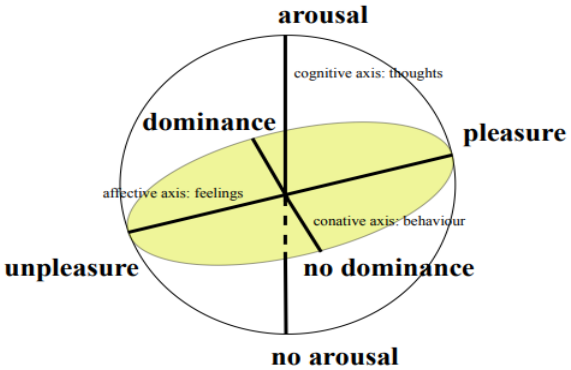

Figure 2: Dimensional Model of PAD as Tripartite View of Experience (Alhasan et al., 2020)

The "pleasure" dimension, often referred to as the "valence" dimension, tells us how pleasant or unpleasant one feels. The "arousal" dimension indicates how active or inactive one feels. Finally, the third dimension "dominance" reports how dominated or in control one feels. For example, anger is an unpleasant emotion of relatively high intensity or arousal and is dominant. Fear, on the other hand, is also unpleasant and of high arousal, but is rather submissive. Boredom, is only a bit unpleasant, of quite a low arousal and primarily of submissive character (Alhasan et al., 2020).

The Role of Deep Learning in SER: Stacked LSTM

Deep learning, especially Stacked Long Short-Term Memory (LSTM) networks, has significantly improved the performance of SER systems. Unlike traditional models like GMMs and SVMs, LSTM networks are specifically designed to handle sequential data, making them ideal for analyzing speech, which has temporal dynamics. Stacked LSTMs consist of multiple layers of LSTM units, allowing the model to capture deeper patterns and long-range dependencies in the data. This is important because emotions in speech are not conveyed in isolated words but through continuous patterns over time. LSTMs excel at remembering important information from earlier points in the sequence while discarding irrelevant data, a key feature for emotion detection. For example, detecting frustration might require understanding the overall tone of a conversation rather than just focusing on a single phrase. The stacked architecture enhances this ability by processing the data through multiple layers, allowing the model to identify both simple and complex emotional cues. By doing so, Stacked LSTM networks significantly improve the accuracy and robustness of SER systems, especially in challenging real-world conditions (Tanoko & Zahra, 2022).

Hidden Markov Models for Temporal Dynamics in Speech Data

Hidden Markov Models (HMM) have been widely used in speech-processing tasks because they are excellent at modeling sequences. In the context of SER, HMMs capture the temporal nature of speech data by breaking it down into smaller, discrete states, each representing a different phase of the speech signal. These states can correspond to different emotional tones or transitions between emotions. By modeling how emotions progress over time, HMMs can provide more contextually accurate predictions compared to static models like SVMs. For instance, consider a person who starts a conversation sounding neutral but gradually becomes angry. An HMM can recognize this progression by estimating the probability of transitioning from a neutral emotional state to an angry one. This ability to model the temporal dynamics of speech makes HMMs well-suited for SER tasks. However, one limitation of HMMs is their reliance on linear assumptions about the transitions between states, which can sometimes oversimplify the complexity of human emotions. This is where combining HMMs with deep learning models like LSTM can offer improved performance (Lin & Busso, 2020). Fig.3. dipicts the Raw waveform plots from EmoDB - using the same sentence and speaker, different portrayed emotion.

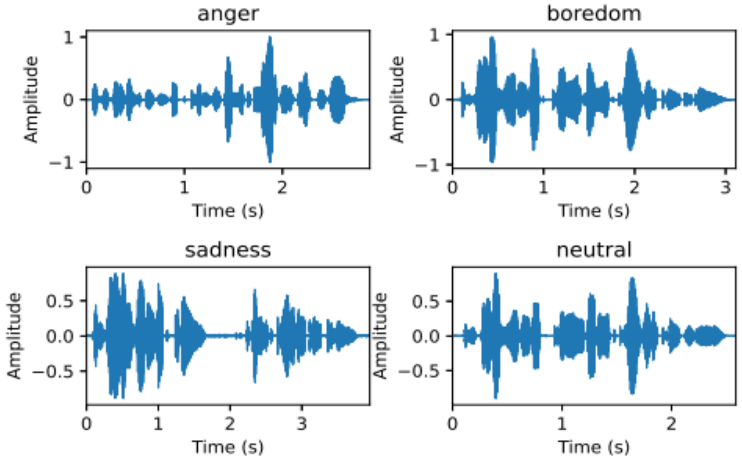

Figure 3: Raw Waveform Plots from EmoDB - using the same Sentence and Speaker, Different Portrayed Emotion

MATERIALS AND METHODS

This paper presents a Stacked Long Short-Term Memory (LSTM) and Hidden Markov Model (HMM) for Speech Emotion Recognition-. The paper examined and emphasized the important role of LSTM and HMM Techniques in Speech, Emotion Recognition. In addition, the research employed the novel preferred reporting items Metha-analysis (PRISMA) in conducting a detail choosy Systematic Review of Existing research. The choice of methodology used, the comprehensive reporting criteria of the studies were carefully studied on LSTM and HMM Speech Emotion recognition to understand the various shortcomings of the past and recent studies presented with potential open areas or research gap for further studies. Afterward, the developed LSTM and HHMM Systematic Related review (SLR) paper focused on investigating the accuracy, efficiency of the various preceding models, what was missing and possible weakness to identify further important problem unattended to on LSTM and HMM hybrid Speech, Emotion Recognition. Part of the findings is reported in Table 1, while subsection A, B, C, and D detailed the PRISMAS Analysis Study providing a comprehensive Summary of various strategies as well as shortcomings of preceding research both past and recent findings on LSTM and HMM paving the way for possible superior research. In the research, a search strategy was developed and utilized to identify unique databases in search of up-to-date, relevant research Articles on LSTM and HMM on Speech, Emotion Recognition.

Database Search Strategy and Eligibility Criteria

In this paper, a search strategy was developed and utilized to strategically identify a variety of databases in search of recent and up-to-date, relevant research publications on LSTM and HMM for speech Emotion recognition. Google Scholar (https://scholar.google.com), ACM Digital Library (https://dl.acm.org/search/), Research Gate (https://www.researchgate.net/lab/), and IEEE Scopus Indexed platforms were all employed in the search. The time frame for the investigation, as published online and referenced on the references list, is Ten years period, ranging

from 2015 to 2024. These sources were selected because of their extensive indexing of research on LSTM and HMM for speech Emotion recognition accessed with less effort. Fig. 4, 5, 6 and 7 depicts the comprehensive systematic analysis of the study while table 1 presents the executive summary.

Review Strategy

This research presents a systemic approach to analyzing research studies. It dissects each study across seven key elements namely: The developed model, the objectives, methodology used, algorithm developed (if applicable), the results Obtained/Findings, The Limitations of the Research/weaknesses, as well as the future work. It then examines the results, identifies knowledge gaps, and suggests future research directions. In addition, the technique considers the research design, search strategy for relevant studies, information sources, and selection process based on preferred reporting Items Metha-Analysis (PRISMA) criteria. The systematic study excluded
commentaries, letters, and editorials to focus on primary research findings. The feedback analysis of the Prisma is depicted in fig. 4.

Search Stage

Search Stage 1 (Information Extraction)

An extensive search was conducted across ten (10) electronic databases. This search yielded 1,025 articles, subsequently subjected to a meticulous selection process.

Search Stage 2 (Screening)

After eliminating duplicate articles and irrelevant works of literature, a residual quantity of items was obtained for subsequent analysis.

Search Stage 3 (Eligibility Determination)

50 articles were selected as relevant to the study while acquiring relevant literature. Following that, some articles were excluded due to the absence of a well-defined methodology and sole reliance on abstracts due to the restriction of paywalls.

Search Stage 4 (Inclusion)

Following the PRISMA research objective, a quality assessment of the remaining articles was performed, and it was determined that only 30 meat the desired criteria for inclusion. Out of the 1,025, 30 finally had the best research methodology with important problems appropriate for inclusion in the subsequent systematic review, as presented in table 1.

Table 1: Executive Summary of the Systematic Review

References	Objectives	Methodology used	Results/findings	Algorithm	Limitation /Weakness	Future Work
(Coto-Jiménez, 2021)	*improve the quality of artificial voices produced using hmm-based statistical parametric speech synthesis. explore the application of discriminative post-filters (using LSTM neural networks) for enhancing synthesized voices.*	Development of discriminative postfilter using LSTM neural networks .Differentiation of voiced and unvoiced segments in speech synthesis, with specific post-filters applied to each segment. Evaluation using five voices, three objective measures, Mel cepstral distance, and subjective tests.	Discriminative postfilter improves the quality of synthesized voices compared to non-discriminative postfilter and standard HTS voices. Objective measures and subjective evaluations confirmed the superiority of the proposed method, with listeners preferring voices processed with discriminative postfilter.	The developed algorithm applies an LSTM-based discriminative postfilter specifically to enhance voiced and unvoiced segments separately rather than applying a single postfilter to the entire speech output.	differentiation between voiced and unvoiced No specific analysis of more granular sound categories like fricatives, plosives, and liquids. the computational complexity of implementing discriminative postfilter in detail was not reported	Analyze specific sound categories like fricatives, plosives, and liquids. Optimizition of computational efficiency for real-time applications. Integration of discriminative postfilter with other advanced speech synthesis frameworks.
(Amami, 2023)	Improve the quality of artificial voices produced using HMM-based statistical parametric speech synthesis. Explore the application of discriminative postfilter, particularly LSTM deep neural networks, for enhancing synthesized voices.	A hybrid deep learning model combining Bidirectional Long Short-Term Memory (BiLSTM) and Convolutional Neural Network (CNN) architectures. Pre-trained CNN is used to extract temporal and spectral features from speech signals, which are input into the BiLSTM network to capture temporal dynamics.	The hybrid BiLSTM-CNN model achieved a high accuracy of 98.86% on a dataset of speech signals from individuals with voice pathologies. This suggests the model's potential as a valuable diagnostic tool for voice pathologies in clinical settings, outperforming several other classifiers.	The developed algorithm is a hybrid BiLSTM-CNN model that combines temporal and spectral features extracted from speech signals to detect different types of voice pathologies.	Limited to evaluation on a single dataset (MEEI database), which may not reflect the variability of real-world clinical settings. Does not explore robustness across different datasets, noise, or recording conditions.	Investigate robustness using different multi-pathologies detectors and additional datasets for broader generalization. Explore the possibility of determining the severity of voice pathology. impact of noise and recording conditions on performance.
(Shahin et al., 2023)	Enhance the quality and	Hybrid model combining Dual-	The DC-LSTM COMP-CapsNet	The developed algorithm is DC-	The model accuracy 57% was	Enhanced performance under

References	Objectives	Methodology used	Results/findings	Algorithm	Limitation /Weakness	Future Work
	efficiency of SER systems using a novel dual-channel LSTM compressed capsule network (DC-LSTM COMP-CapsNet). Improve compression in CapsNet for SER. Develop a text-independent and speaker-independent SER model.	Channel LSTM (DC-LSTM) and Compressed Capsule Networks (CapsNet). Mel-frequency cepstral coefficients (MFCCs) delta-delta feature extraction. Grid search (GS) for optimizing model parameters. Evaluation on four speech datasets: Arabic Emirati-accented corpus, SUSAS, RAVDESS, and CREMA-D.	achieved an average emotion recognition accuracy of 89.3% on the Arabic Emirati-accented corpus, surpassing traditional classifiers and other state-of-the-art systems like CapsNet (84.7%), CNN (82.2%), SVM (69.8%), MLP (69.2%), and KNN (53.8%). Improved performance, training, and testing times were also noted.	LSTM COMP-CapsNet, which integrates dual-channel LSTM layers to capture sequence-correlated characteristics of speech signals, followed by compressed capsule networks for emotion recognition.	not optimal due to dataset outliers. CapsNet can misinterpret voice complexity. 3. Higher feature extraction and training times compared to conventional classifiers. The study does not address robustness across datasets or noise conditions.	emotional and stressful conditions for speaker identification and verification. Reduce computational complexity and training/testing times. Explore distributed DCT based on MFCC to improve model robustness and efficiency.
(Patel et al., 2024)	Provide a comprehensive overview of Speech Emotion Recognition (SER) methods and datasets from recent years. Identify research gaps and challenges in SER. Compare different SER techniques and datasets. focusing on transformer-based and multimodal data integration.	Systematic review and analysis of literature on SER methods and datasets. Comparative analysis using attention mechanisms to evaluate the effectiveness of machine learning, deep learning, and transformer-based techniques for SER.	The survey found that deep learning and transformer-based approaches, especially those incorporating attention mechanisms, outperform traditional SER methods. The study highlights key challenges in SER and suggests future research directions to enhance system performance.	No new algorithm is proposed; instead, the study synthesizes and compares existing SER methods, emphasizing the superior performance of deep learning and transformer-based approaches with attention mechanisms.	Limited discussion on real-world application challenges (e.g., noise, environmental factors). The rapidly evolving field may limit the survey's coverage of the latest advancements.	Further exploration of transformer-based SER techniques. Integration of real-time speech data with facial expressions to enhance multimodal emotional expression detection. Addressing the challenges in practical SER applications with more complex and accurate systems.
(Abbaschian et al., 2021)	analyse various deep learning techniques used in SER, such as DNNs, LSTMs, and attention mechanisms. Provide a multi-aspect comparison of neural network approaches in SER. Highlight the strengths, limitations, and future directions for SER.	a comprehensive review of SER databases (e.g., EMO-DB, DES, IEMOCAP, VAM) and deep learning techniques (DNNs, LSTMs, attention mechanisms). The study conducts a multi-aspect comparison of neural network approaches in SER, focusing on factors such as sample duration, citation frequency, and database challenges.	CNNs show strong low-level and short-term discriminative capabilities. Deep convolutional LSTM structures enhance long-term memory and speaker-independent emotion recognition. Attention mechanisms improve system efficiency by adding nonlinearity. Larger databases have an average sample duration of 2.8 seconds.	The study reviews existing algorithms without proposing a new one. It focuses on evaluating the effectiveness of deep learning techniques (DNNs, LSTMs, attention mechanisms) for SER.	Over-reliance on accuracy as a performance measure, lacking comprehensive metrics like precision, recall, and F1-score. Variations in database sample duration and characteristics challenge uniform evaluation.	Develop more robust, dataset-independent SER models. Improve performance metrics beyond accuracy to include precision, recall, and F1-score. 3. Focus on building production-ready models that apply to real-world use cases and diverse datasets.
(Ullah et al., 2023)	Address the challenge of effectively extracting emotional features from speech utterances.	The model parallelizes two CNNs for spatial feature extraction and a Transformer encoder for temporal feature extraction. The	The CTENet model achieves 82.31% accuracy on the RAVDESS dataset (eight emotions) and 79.42% accuracy on the IEMOCAP dataset	CTENet (CNN-Transformer Encoder Network), which fuses CNNs for spatial feature extraction and a Transformer	Focuses primarily on accuracy, lacking other performance metrics like precision, recall, and F1-score.	Increase the number of database entries for improved accuracy. Refine the model architecture and explore better feature extractors.

References	Objectives	Methodology used	Results/findings	Algorithm	Limitation /Weakness	Future Work
	Combine spatial and temporal feature representations for better SER. Achieve higher accuracy in recognizing speech emotions using a fusion model (CNN + Transformer encoder).	RAVDESS dataset is used for recognizing eight emotions, while the IEMOCAP dataset evaluates the model for five emotions. Additive White Gaussian Noise (AWGN) is applied for dataset augmentation.	(five emotions). The model outperforms state-of-the-art SER models, showing effectiveness in combining spatial and temporal features. It is computationally efficient and compact (4.54 MB).	encoder for temporal feature extraction from the MFCC spectrum treated as grayscale images, offering hierarchical featurerepresentation at a lower computational cost.	The model overfits when using fewer dataset entries, although dropout regularization can help. Needs more robust feature sets and modalities to improve accuracy.	Combine different feature sets and modalities for more robust training and higher accuracy. Apply newer models to achieve state-of-the-art SER results.
(Haque et al., 2023)	Overcome the shortcomings of traditional SER approaches in capturing long-term dependencies, temporal dynamics, and complex patterns in multimodal data. Develop an ensemble model leveraging Graph Convolutional Networks (GCNs) and HuBERT transformer for improved SER. Enhance accuracy in emotion recognition by fusing textual and audio modalities.	Graph Convolutional Networks (GCNs) process textual data to capture long-term contextual dependencies and relationships using graph-based representations. Self-attention mechanisms are used to capture long-range dependencies and model temporal dynamics in speech data. Combines GCN and HuBERT for simultaneous multimodal data analysis.	The ensemble model combining GCN and HuBERT shows enhanced accuracy in recognizing emotions from speech compared to traditional methods. It successfully leverages the strengths of GCNs for textual data and HuBERT for audio data, leading to better discriminative power in SER.	The ensemble model combines Graph Convolutional Networks (GCNs) to process textual features and the HuBERT transformer to analyze audio signals. This fusion enables the model to capture local, global, and long-range dependencies in speech data, improving emotion recognition accuracy.	Lacks comprehensive performance metrics like precision, recall, and F1-score. Does not discuss computational efficiency or real-world applicability in detail. Potential overfitting with limited dataset entries, despite dropout regularization. Limited discussion on computational resource requirements and scalability.	Refine the model architecture to improve accuracy further. Combine different feature sets for more robust training. Add modalities such as visual cues to increase recognition accuracy. Apply newer models and conduct evaluations using broader performance metrics. Explore scalability for real-world applications.
(Abbaschian et al., 2021)	Analysed deep learning techniques (DNNs, LSTMs, attention mechanisms). Provide multi-aspect comparison of neural network approaches in SER. Highlight the strengths and limitations of current methods. Suggest future research directions.	Reviewed databases: EMO-DB, DES, IEMOCAP, VAM, and newer English databases. Analyzed major deep learning techniques: DNNs, LSTMs, and attention mechanisms. Conducted multi-aspect comparison focusing on sample duration, citation, etc.	Newer databases have larger samples CNNs show better low-level and short-term discriminative capabilities. Deep convolutional LSTMs improve long-term memory and speaker independence. Attentionmechanisms add nonlinearity and efficiency. Over-reliance on accuracy as the primary performance measure.	DNNs (Deep Neural Networks) and LSTMs (Long Short-Term Memory Networks) Attention Mechanisms. CNN (Convolutional Neural Networks) Deep convolutional LSTM structures	Over-reliance on accuracy as the primary performance measure. Limited practical implementation and production-readiness of models. Inconsistencies in sample durations across databases. Lack of dataset independent solutions for real-world applications.	Develop robust, dataset-independent SER solutions. Improve performance measures beyond accuracy (precision, recall, F1-score). Focus on creating production-ready models for diverse datasets.
(Huang et al., 2021)	improve continuous emotion recognition by incorporating long-term temporal	Utilized a window of feature frames, applied frame skipping and temporal pooling at the feature level, and skip RNN at	Skip LSTM outperforms standard LSTM models by focusing on critical emotional states and skipping trivial	Skip Recurrent Neural Network (Skip RNN) with Skip Long Short-Term Memory (Skip LSTM).	Redundancy in emotional features must be addressed, and comprehensive measures beyond accuracy are	Enhance the performance of Skip LSTM and explore multimodal emotion fusion to improve temporal modeling for

References	Objectives	Methodology used	Results/findings	Algorithm	Limitation /Weakness	Future Work
	context, using skip RNN and methods like frame skipping and temporal pooling to reduce redundancy and enhance performance.	the model level to capture long-term emotional context. Experimental evaluation was conducted using the AVEC 2017 database.	information. The longer window enhances the ability to model temporal emotional contexts, improving continuous emotion recognition accuracy.		required to better evaluate models.	real-world applications.
(Venkatara manan & Rajamohan, 2019)	To compare various speech-based emotion recognition approaches using different feature extraction techniques and machine learning models, and determine the most effective combinations for emotion classification.	Audio recordings from the RAVDESS dataset were pre-processed, and features such as Log-Mel Spectrogram, MFCCs, pitch, and energy were extracted. Models like LSTM, CNNs (2D & 3D), HMM, and DNN were applied and evaluated for classification accuracy.	The 2D CNN with Log-Mel Spectrogram features achieved the best accuracy of 68% in the 14-class emotion classification task. Gender-specific classification improved performance due to pitch and energy differences between male and female voices.	2D Convolutional Neural Network (CNN) with Log-Mel Spectrogram features.	Limited dataset size led to lower accuracy; the subjective nature of human emotion perception complicates the problem, making it difficult to achieve higher accuracy across different emotions.	Further exploration of robust feature engineering techniques and larger datasets to improve emotion recognition accuracy, and extending the models to work across multiple languages and diverse emotional contexts.
(Mohamed & Schuller, 2020)	To develop Conceal Net, an end-to-end recurrent neural network for real-time packet loss concealment and speech emotion recognition in environments with frequent packet losses.	The study used stacked generative RNN cells (LSTM layers) wrapped with a concealment cell for audio reconstruction. ConcealNet was evaluated using the RECOLA dataset and compared against baselines (0-substitution and linear interpolation). Stress training was applied to handle long-term packet losses.	ConcealNet showed significant improvement in audio reconstruction and emotion prediction, especially in scenarios with short packet losses. For arousal, CCC dropped slightly from 76.93% to 75.99%, and for valence, from 43.18% to 39.81%. The bidirectional variant performed even better.	ConcealNet, a stacked RNN with LSTM layers and a concealment wrapper for end-to-end packet loss concealment and emotion prediction.	ConcealNet struggles with long-duration packet losses, despite the introduction of stress training to mitigate this issue.	Future work could explore attention mechanisms and generative approaches, including GANs or variational models, to improve packet loss concealment and emotion recognition performance.
(Tzinis & Potamianos, 2017)	investigate the performance of RNN models in Speech Emotion Recognition (SER) by analyzing both local and global features at various time-scales (frame, phoneme, word, or utterance).	The study evaluated RNN performance using Low-Level Descriptors (LLDs) and statistical functionals across different time-scales. Experiments were conducted using the IEMOCAP corpus, and the optimal time-scale was found at the word level, using an LSTM model.	The study achieved state-of-the-art SER performance on the IEMOCAP corpus by extracting statistical features over speech segmentscorrespon ding to a couple of words. This method also reduced model and computational complexity compared to previous approaches.	LSTM-based RNN model that extracts statistical features over speech segments at a word-level time-scale for SER.	The model was only tested on the IEMOCAP corpus, and further validation on other datasets is required. The study also lacks a mechanism for multi-scale decision-making.	Future work aims to introduce a multi-scale decision-making mechanism for emotional sequences and evaluate the model on additional databases.
(Ahmed et al., 2023)	To propose an ensemble model combining CNN, LSTM, and GRU	Three architectures were used: CNN-FCN, LSTM-FCN, and GRU-FCN. Data augmentation	The ensemble model achieved state-of-the-art accuracy: 99.46% (TESS), 95.42% (EMO-DB), 95.62%	Ensemble of CNN-LSTM-GRU model using weighted average of three architectures	Limited by small, acted datasets, making models prone to overfitting. Real-world noisy	Investigating multi-label emotion recognition in continuous speech, integrating attention mechanisms, and

References	Objectives	Methodology used	Results/findings	Algorithm	Limitation /Weakness	Future Work
	architectures for speech emotion recognition, improving accuracy by leveraging local and global features.	(e.g., white Gaussian noise, pitch shifting) was applied to five datasets (TESS, EMO-DB, RAVDESS, SAVEE, CREMA-D).	(RAVDESS), 93.22% (SAVEE), and 90.47% (CREMA-D), outperforming other models.	(CNN-FCN, LSTM-FCN, GRU-FCN), with handcrafted features from speech signals.	environments were not addressed, and further improvements are needed for robustness.	exploring phase-based acoustic features for better SER performance.
(Jalal et al., 2019)	To propose a novel hybrid architecture combining BLSTM, CNN, and Capsule Networks for robust emotion classification from speech signals.	A hybrid framework of BLSTM for sequential speech data processing, CNN for feature extraction, and Capsule Networks for clustering. Tested on FAU-Aibo and RAVDESS datasets.	Achieved state-of-the-art accuracy: 77.6% on FAU-Aibo and 56.2% on RAVDESS, with improvements of 3% and 14% over previous best results for respective tasks.	BLSTM combined with 1D Conv-Capsule layers, leveraging Capsule routing for hierarchical temporal modeling and clustering of emotion-related speech features.	evaluation on only two datasets (FAU-Aibo and RAVDESS). Generalizability to other speech-based tasks like speaker or language recognition is unexplored.	Expanding the application of the proposed architecture to tasks such as speaker and language recognition, and further testing on diverse datasets.
(Tanoko & Zahra, 2022)	To analyze the impact of the stacking order of multiple speech features on the performance of speech emotion recognition (SER) using 1D CNN architecture.	Brute force approach to test all possible combinations of five features (MFCC, Mel-spectrogram, chromagram, spectral contrast, tonnetz) and evaluate SER performance on the RAVDESS dataset using 1D CNN.	The best combination achieved 79.17% accuracy for 8 emotion classes. Changing the feature order improved accuracy by 16.05% and reduced model size by 96%.	1D Convolutional Neural Network (CNN) model with different orders of stacked features (spectral contrast, tonnetz, chromagram, Mel-spectrogram, and MFCC).	The study is limited to five specific features and the RAVDESS dataset; generalizability to other features or datasets is not tested.	Investigating additional features like Teager energy and applying the approach to other datasets to further improve performance and robustness.
(Lin & Busso, 2020)	To develop an efficient temporal modeling approach for speech emotion recognition (SER) by splitting varied duration sentences into a fixed number of chunks for better feature representation and processing efficiency.	The approach segments speech data into fixed-duration chunks by varying the overlap between chunks, enabling temporal modeling with long short-term memory (LSTM) and attention mechanisms. The methodology was evaluated using the MSP-Podcast dataset.	The proposed method showed significant improvements in recognition accuracy and computational efficiency compared to alternative temporal-based models that rely on LSTM.	Chunk-level temporal modeling combined with LSTM networks and attention mechanisms for speech emotion recognition.	The approach was only tested on a single dataset (MSP-Podcast), limiting its generalizability to other datasets and sequence-to-one tasks.	Future research includes validating the approach on multiple datasets, exploring CNN and DNN for feature extraction, and analysing chunk-level attention weights for emotion externalization.
(Sönmez & Varol, 2020)	To develop a lightweight, effective Speech Emotion Recognition (SER) method with low computational complexity using a combination of texture analysis techniques.	The method applies a one-dimensional discrete wavelet transform (DWT) on raw audio data, extracts features using local binary pattern (LBP) and local ternary pattern (LTP), and selects 1024 dominant features using Neighbourhood Component Analysis (NCA)	The model, called 1BTPDN, achieved high recognition rates across different databases: 95.16% (RAVDESS), 89.16% (EMO-DB), 76.67% (SAVEE), and 74.31% (EMOVO), outperforming many state-of-the-art SER methods.	The 1BTPDN algorithm combines 1D-LBP, 1D-LTP, DWT, and NCA to extract key features, with classification done via a polynomial kernel-based SVM.	The model was tested on offline databases, lacking real-time audio data validation. The model's generalization across subjects and sentences needs further exploration through cross-validation.	Future studies will explore real-time audio datasets, cross-subject, and cross-sentence validation for better generalization. Additionally, more effective methods can be developed for new datasets.

References	Objectives	Methodology used	Results/findings	Algorithm	Limitation /Weakness	Future Work
		for classification by a third-degree polynomial kernel-based Support Vector Machine (SVM).				
(Zheng et al., 2019)	To improve the accuracy and generalization of speech emotion recognition (SER) by designing a multi-level ensemble learning model that extracts global and local emotion features.	(CNN, BLSTM, and GRU) designed to extract speech features, focusing on local signals. An BLSTM, ensemble learning model was developed to combine their strengths using a CRNN-based internal training mechanism with attention mechanisms. The model was evaluated using the IEMOCAP corpus.	The ensemble model improved speech emotion recognition accuracy and generalization by combining multi-level global and local features. It addressed the data imbalance problem and achieved better performance compared to individual models.	The ensemble learning model combines CNN, BLSTM, GRU, and CRNN (with attention mechanisms) to extract multi-level features and perform SER.	The proposed model was tested on the IEMOCAP dataset, which may limit its performance in real-world applications due to dataset-specific biases.	Future research will focus on developing a personalized network model that combines both general and speaker-specific features to further improve the efficiency and accuracy of SER.
(Zhang et al., 2021)	To develop a Heterogeneous Parallel Convolution Bi-LSTM (HPCB) model to improve the accuracy and effectiveness of speech emotion recognition (SER) by exploiting spatiotemporal information.	The HPCB model utilizes two parallel branches: one with two dense layers and a Bi-LSTM layer, and the other with a dense layer, a convolution layer, and a Bi-LSTM layer. The model was trained and tested on three databases: EMO-DB, CASIA, and SAVEE.	The HPCB model achieved unweighted average recalls of 84.65% on EMO-DB, 79.67% on CASIA, and 56.50% on SAVEE, demonstrating better performance compared to prior models.	The Heterogeneous Parallel Convolution Bi-LSTM (HPCB) model with two branches: one with dense layers and Bi-LSTM, and another with convolution and Bi-LSTM layers for enhanced spatiotemporal feature extraction.	Potential overfitting due to the complexity of the model architecture, despite using a 0.5 dropout ratio.	Further verification on additional databases and comparison with other deep learning models (e.g., GANs, zero-shot learning) to reduce overfitting risks and expand to other domains like audio and image recognition.
(Wu et al., 2019)	To develop a novel architecture using Capsule Networks (CapsNets) for improving Speech recognition (SER) by capturing the spatial relationships in speech features.	The proposed method employs CapsNets, which capture spatial relationships in speech features like pitch and formant frequencies. Recurrent connections are introduced to improve time sensitivity, and experiments are conducted on the IEMOCAP dataset.	The CapsNet-based model outperformed the baseline CNN-LSTM model, achieving 72.73% weighted accuracy (WA) and 59.71% unweighted accuracy (UA) on the IEMOCAP dataset.	Capsule Networks (CapsNets) with recurrent connections, combined with CNN and GRU layers, referred to as the CNN-GRU-SeqCap architecture, to enhance feature extraction and classification.	The paper highlights the need for further improvement in capturing time-varying emotional characteristics, which is a limitation of the current model.	Future research will focus on extending CapsNets' ability to capture emotions that vary over time and applying the architecture to more diverse emotional datasets.
(Le et al., 2017)	To develop a novel approach for continuous speech emotion recognition by discretizing training labels and employing a multi-task deep bidirectional	A multi-task BLSTM recurrent neural network is trained with cost-sensitive cross-entropy loss to model discretized emotion labels, incorporating an emotion decoding algorithm for robust time series estimates.	The proposed method achieved competitive performance on the RECOLA dataset, outperforming previous works and strong regression baselines in audio-only emotion recognition tasks.	Multi-task deep bidirectional long-short-term memory (BLSTM) recurrent neural network, combined with an emotion decoding algorithm that leverages long- and short-term signal properties.	The paper does not extensively address potential limitations of the approach or variations in emotional expression across different speakers.	Future work may focus on exploring the application of this classification-based approach to other emotion recognition tasks and refining the model to enhance performance across diverse datasets.

References	Objectives	Methodology used	Results/findings	Algorithm	Limitation /Weakness	Future Work
	long-short-term memory (BLSTM) network.					
(Lee & Tashev, 2015)	To develop a speech emotion recognition system using a recurrent neural network (RNN) model that captures long-range context effects and addresses the uncertainty of emotional label expressions.	The system employs a bidirectional long short-term memory (BLSTM) model for high-level representation of emotional states, incorporating an efficient learning algorithm to train sequences of random variables.	The proposed system achieved a weighted accuracy improvement of up to 12% compared to a baseline emotion recognition system based on DNN-ELM.	The algorithm combines a recurrent neural network approach with maximum-likelihood-based learning to effectively model the long-range context of emotional speech.	The paper does not explicitly address potential challenges related to the variability of emotional expression across different speakers or environmental factors.	Future research could explore further refinements in the model to enhance its robustness and applicability across diverse speech emotion recognition tasks and datasets.
(Nakisa et al., 2018)	To propose a framework for the automatic optimization of LSTM hyperparameters using differential evolution (DE) to improve emotion classification accuracy.	The framework employs the DE algorithm for hyperparameter optimization, comparing its performance against other methods like particle swarm optimization (PSO) and simulated annealing (SA) on data collected from wearable sensors.	The optimized LSTM model achieved an accuracy of 77.68% for four-quadrant dimensional emotions, with a 14% increase in accuracy compared to baseline models.	The DE algorithm is used for optimizing LSTM hyperparameters, particularly batch size and number of hidden neurons, to enhance emotion recognition performance.	The DE algorithm is computationally expensive, and the study notes potential issues with premature convergence, which can trap the optimization process in local optima.	Future research should explore new evolutionary computation algorithms to address premature convergence and improve emotion classification performance, as well as leveraging parallel or cloud computing to reduce processing time.
(Chamishka et al., 2022)	To propose a novel feature extraction technique using Bag-of-Audio-Words (BoAW) and develop a Recurrent Neural Network (RNN) model for real-time emotion detection from audio data, aiming to improve the accuracy of emotion classification compared to existing methods.	The study employs BoAW for feature extraction and RNN for emotion detection, evaluating its performance on benchmark datasets, including the IEMOCAP dataset, through empirical testing for real-time emotion prediction capability.	The proposed method achieved 60.87% weighted accuracy and 60.97% unweighted accuracy in detecting six basic emotions on the IEMOCAP dataset, outperforming current state-of-the-art models by approximately 20%.	A Recurrent Neural Network (RNN) model for emotion detection, enhanced with Bag-of-Audio-Words (BoAW) feature embeddings, enabling the system to capture contextual conversation states.	The main limitation identified is the lack of robust real-time automatic speaker diarization, which hinders the deployment of the emotion detection model in real-world applications.	Future research directions include extending the model to detect mixed emotions, integrating textual features from Bag-of-Words (BoW), exploring deep learning techniques like wave2vec 2.0 for feature extraction, and developing a fully automated emotion recognition pipeline.
(Zhao et al., 2019)	To develop and evaluate 1D and 2D CNN LSTM networks for learning deep emotion features to improve speech emotion recognition accuracy from raw audio	Two CNN LSTM networks (1D and 2D) were constructed, each with four local feature learning blocks (LFLBs) and one LSTM layer to learn local and global emotion-related features from audio data.	The 2D CNN LSTM network achieved recognition accuracies of 95.33% (speaker-dependent) and 95.89% (speaker-independent) on the Berlin EmoDB and outperformed traditional methods on the IEMOCAP database.	The networks utilized a combination of convolutional layers and LSTM layers for feature extraction, where LFLBs learned local features and the LSTM layer captured long-term dependencies.	The "black box" nature of the deep networks remains unexplained, making it difficult to interpret how emotions are recognized, and further research is needed to enhance understanding of these networks.	Future research could focus on improving emotion recognition accuracy, uncovering the workings of the deep networks, developing new architectures, and merging features learned from different deep networks.

References	Objectives	Methodology used	Results/findings	Algorithm	Limitation /Weakness	Future Work
	clips and log-mel spectrograms.					
(Wu et al., 2021)	To develop a novel sequential capsule network (CNN_SeqCap and CNN_RecCap) for speech emotion recognition (SER) that can preserve spatial information from spectrograms and capture temporal information to improve emotion prediction accuracy.	The study introduces two new architectures: sequential capsule networks (CNN_SeqCap) and recurrent capsule networks (CNN_RecCap). Dynamic routing algorithms are used for obtaining utterance-level features, with evaluations performed using the IEMOCAP dataset.	CNN_SeqCap improves weighted accuracy (WA) by 3.76% and unweighted accuracy (UA) by 0.11%. CNN_RecCap improves WA by 4.52% and UA by 1.57% over the CNN baseline.	Sequential capsule networks (CapNets) use dynamic routing to preserve spatial information from spectrograms. Recurrent connections are introduced to capture temporal information, enabling better performance in recognizing emotions in speech.	The recurrent connection does not capture long-distance temporal dependencies well enough, limiting the system's overall ability to recognize emotions in speech, especially for long utterances.	The researchers plan to enhance the recurrent connection to capture longer-distance context to improve the recognition performance of the sequential capsule networks.
(Lim et al., 2021)	To propose a Speech Emotion Recognition (SER) method based on a combination of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) without using traditional hand-crafted features, aiming to improve classification accuracy.	The study employs a concatenated architecture of CNNs and RNNs (with LSTM layers) for SER, extracting features directly from an emotional speech database without traditional hand-crafted features. The CNN extracts hierarchical features, and the RNN captures sequential information.	The CNN-RNN hybrid approach outperforms conventional classification methods in recognizing emotions from speech. Time-distributed CNNs show even better performance than basic CNNs and LSTM-based models, setting a baseline for future research.	A CNN-RNN concatenated model where CNNs perform hierarchical feature extraction from speech spectrograms, followed by LSTM layers to handle temporal information. Time-distributed CNNs are introduced to improve performance.	The model is limited to audio-based emotion recognition and does not consider multimodal data, such as visual cues. Additionally, the study doesn't utilize hand-crafted features, might miss relevant for emotion classification.	Future plans include expanding the model to multimodal emotion recognition tasks, combining audio and video inputs to further improve the emotion recognition performance.
(Mu et al., 2017)	Propose a method for Speech Emotion Recognition (SER) that leverages distributed Convolutional Neural Networks (CNNs) and Bidirectional Recurrent Neural Networks (BRNNs) with an attention mechanism to improve accuracy and interpretability in identifying	The approach involves using CNNs to automatically learn features from raw spectral information, followed by BRNNs to capture temporal information, and implementing an attention mechanism to focus on emotion-relevant parts of the utterance.	The proposed method achieved 64.08% weighted accuracy and 56.41% unweighted accuracy for four-emotion classification on the IEMOCAP dataset, outperforming previous results reported for the same dataset, demonstrating the effectiveness of the attention mechanism.	A hybrid architecture combining distributed CNNs for feature extraction and BRNNs for temporal analysis, enhanced by an attention mechanism that focuses on relevant segments of the output sequence to improve classification accuracy and model interpretability.	The study is limited to four-emotion classification and does not consider multimodal inputs, which could enhance emotion detection. The focus on specific emotions may also restrict the generalizability of the results to broader emotional contexts.	Future research plans include extending the approach to multimodal emotion recognition and transitioning from a four-emotion classification task to a regression prediction model for arousal, valence, and liking, aiming for more comprehensive and accurate emotion detection.

References	Objectives	Methodology used	Results/findings	Algorithm	Limitation /Weakness	Future Work
	emotions from speech.					
(Harby et al., 2024)	To enhance the accuracy of Speech Emotion Recognition (SER) by systematically selecting multiple audio cues and utilizing deep Bi-LSTM models for effective emotion discernment.	The study utilized a 2D Convolutional Neural Network (CNN) for preprocessing audio signals, followed by sequential feature selection using Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS), and deep Bi-LSTM for emotion classification.	The proposed SER model achieved accuracy rates of 90.92% for IEMOCAP, 93% for EMO-DB, and 92% for RAVDESS datasets, indicating improved performance compared to state-of-the-art SER techniques.	The algorithm involves feature extraction from spectrograms using CNN, followed by feature selection techniques (SFS and SBS), and classification using a deep Bi-LSTM network for emotion recognition.	The use of MFCCs may not capture dynamic changes in emotional expression over longer time scales and primarily focuses on acoustic features, missing semantic content representation.	Future research could explore longer analysis windows, both topologies of deep Bi-LSTM in the framework, and integration of audio data with visual modalities to improve emotion recognition further.
(Van Houdt et al., 2020)	To provide a comprehensive review of Long Short-Term Memory (LSTM) models, covering their formulation, training, and applications across various domains such as time series forecasting, natural language processing, and computer vision.	The paper reviews recent literature and applications of LSTM, with a focus on understanding how LSTM handles the vanishing/exploding gradient problem. It also explores the hybridization of LSTM with CNNs for performance improvement.	LSTM's ability to handle a wide range of problems, such as time series forecasting and sentiment analysis, is well-established. Hybrid models (CNN-LSTM) reduce dimensionality and improve performance in many applications. However, no single variant outperforms standard LSTM across all tasks.	The standard LSTM architecture, often integrated with CNNs for hybrid models, is employed for applications like time series forecasting, image captioning, and natural language processing.	LSTM does not universally outperform other models in all aspects, and improvements are still needed in hybrid architectures. There are limitations in recommendations as results vary based on the heterogeneity of problems.	Future work could focus on customizing hybrid architectures (e.g., CNN-LSTM) to further enhance performance, exploring new integration techniques, and improving LSTM models for specific applications such as image captioning and time series analysis.

Hybrid Models: Integrating LSTM and HMM for SER

Combining LSTM and HMM into a hybrid model leverages the strengths of both approaches, making it possible to create more accurate and context-aware SER systems. While LSTMs excel at capturing long-range dependencies in sequential data, HMMs are good at modelling the step-by-step transitions between different emotional states. The hybrid model allows the system to track not only the temporal dependencies in the data but also the specific transitions between emotional states. In practical terms, the LSTM component of the hybrid model can first analyse the speech signal to extract meaningful patterns and emotional cues over time. The HMM then takes over to model the progression of these emotional cues, making it easier to predict complex emotional shifts, such as transitioning from sadness to frustration. Integrating LSTM and HMM enhances the model's ability to deal with real-world speech, where emotions are rarely static and often change throughout the conversation (Lin et al., 2024)(Gendron & Guibon, 2024) his combination significantly boosts the model's ability to detect subtle emotional nuances, making it particularly useful in applications like mental health monitoring and customer service (Sönmez & Varol, 2020).

Preprocessing Techniques for Speech Emotion Recognition

Preprocessing is a crucial step in any SER system because raw speech data often contains noise and irrelevant information that can degrade model performance. Common preprocessing techniques include Mel-Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding Coefficients (LPCC) (Wang et al., 2025)(Zhao et al., 2025) hese features capture essential properties of the speech signal, such as its frequency and amplitude, while filtering out irrelevant noise. MFCC is perhaps the most widely used technique (Chen & Zhang, 2024), as it mimics how the human ear perceives sound, emphasizing the frequencies most relevant to human speech. LPCC, on the other hand, provides a mathematical model of the vocal tract, which can be used to capture the distinctive qualities of a speaker's voice. By using these preprocessing methods, SER systems can focus on the most informative aspects of the speech signal, improving both speed and accuracy. Other techniques, like voice activity detection (VAD) and noise reduction, are also used to clean up the audio data, ensuring that the models are only analysing the relevant portions of the speech signal (Wang et al., 2025)(Zhao et al., 2025).

Comparative Performance of LSTM, HMM, and Hybrid Models in SER

In (Nam & Park, 2024) Their multi-decoder Wave-U-Net SER model achieves 66.2% accuracy on clean speech (∞ SNR) and 62.4% accuracy at 0 dB SNR, representing a degradation of only 3.8% compared to clean conditions.

[36]The TRNet model reports that it "substantially increases the system's robustness in both matched and unmatched noisy environments, without compromising its performance in clean environments."

Although the arXiv version does not provide a single "accuracy drop %" like the Wave-U-Net paper, the authors emphasize that TRNet maintains high performance even when the noise type during inference does not match the training noise.

Comparison: The Wave-U-Net model gives very concrete numeric robustness metrics (percent-accuracy at different SNRs), demonstrating only modest degradation at high noise. TRNet, on the other hand, emphasizes generalization to unseen noise types and shows robust performance in both matched/unmatched noise, though without the same detailed SNR-by-SNR accuracy breakdown.

When comparing the performance of LSTM, HMM, and hybrid models in Speech Emotion Recognition, studies (Chen & Zhang, 2024)(Nam & Park, 2024) generally show that hybrid models perform better in handling real-world data. LSTM networks outperform HMMs when it comes to capturing long-term dependencies in speech, making them highly effective in recognizing emotions that span multiple sentences or phrases. However, HMMs excel at modelling the transitions between emotional states (Gendron & Guibon, 2024)(Zhang et al., 2024), which is something LSTMs may struggle with in isolation. Thus, Hybrid models, combining both LSTM and HMM, offer the best of both worlds. In addition, they can handle long sequences of speech while also tracking the emotional shifts that occur throughout the conversation. Several research studies comparing these models on benchmark datasets like IEMOCAP and RAVDESS consistently demonstrate that hybrid models achieve higher accuracy rates, particularly in noisy and complex environments. The ability to model both long-term dependencies and short-term transitions makes hybrid models the most promising approach for SER in real-world applications (Zheng et al., 2019).

Benchmark Datasets for Speech Emotion Recognition

Benchmark datasets are essential for training and evaluating SER models. Some of the most commonly used datasets in SER research include IEMOCAP, RAVDESS, and CREMA-D. IEMOCAP (Interactive Emotional Dyadic Motion Capture Database) is one of the most widely used datasets, containing over 12 hours of speech data with emotions like happiness, anger, and sadness labeled by human annotators. It is highly valued for its diverse range of emotions and real-world conversational data, making it ideal for training robust SER models.

RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) is another popular dataset that provides not only speech but also song recordings with emotional labels. This dataset is particularly useful for applications in the entertainment industry, where detecting emotions in music or voice acting is important. CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset) focuses on a wider range of emotions and includes both visual and audio cues, allowing researchers to explore multimodal emotion recognition systems. These benchmark datasets are critical for evaluating new models and ensuring that they perform well in a variety of real-world scenarios.

Evaluation Metrics in Speech Emotion Recognition Models

To measure the effectiveness of Speech Emotion Recognition models, various evaluation metrics are used, with accuracy, F1-score, and the confusion matrix being the most common (Coto-Jiménez, 2021)(Lin et al., 2024). Accuracy is the simplest metric, measuring the percentage of correctly predicted emotions out of the total predictions. However, accuracy alone is not enough, especially in cases where the dataset is imbalanced—where one emotion is more represented than others.

In such cases, the F1-score is a more balanced metric, as it takes both precision and recall into account. Precision refers to the proportion of true positive predictions out of all positive predictions, while recall refers to the proportion of true positives out of all

actual positives. The F1 score harmonizes these two values, providing a better sense of the model's performance in handling different emotions. The confusion matrix is another valuable tool, offering insight into which emotions are being confused with others, and helping developers identify areas for improvement in the model.

Addressing Bias and Fairness in Speech Emotion Recognition

Context-aware emotion recognition systems improve accuracy by incorporating additional information such as conversational history, speaker role, environmental cues, and temporal context rather than relying on raw speech or facial features alone (Gendron & Guibon, 2024). Recent models such as the Self Context-Aware Model (SCAM) use multimodal representations and continuous context tracking to enhance human robot interaction emotion understanding. Other approaches, like the SEC metric-learning framework, embed conversational dependencies to improve emotion recognition in dialogue settings (Gendron & Guibon, 2024). These models typically rely on architectures such as transformers, graph attention networks, and metric-learning Siamese networks, which are better at capturing long-range dependencies and emotion shifts across conversations (Lin et al., 2024).

Bias and fairness are critical concerns in Speech Emotion Recognition systems, particularly because emotions can be influenced by a person's cultural background, gender, or age. For instance, a model trained on data from a particular cultural group might misinterpret emotions from speakers of another group. Similarly, emotions expressed by women may be interpreted differently from those expressed by men, leading to biased outcomes. To address these issues, researchers are focusing on creating diverse and representative datasets that cover a wide range of demographic groups. Additionally, fairness algorithms are being integrated into SER models to ensure that they do not favor one group over another. Regular audits and bias detection tools are also being used to monitor model performance and ensure fairness. This is particularly important in applications like mental health monitoring or law enforcement, where biased outcomes could have serious consequences.

RESULTS AND DISCUSSION

In this systematic review, we explored the advancements in Speech Emotion Recognition (SER), focusing on the hybrid approach that combines Stacked Long Short-Term Memory (LSTM) networks with Hidden Markov Models (HMM). Traditional models such as Gaussian Mixture Models (GMM) and Support Vector Machines (SVM) have been instrumental in laying the groundwork for SER, but their static nature limits their effectiveness in capturing the temporal dynamics of speech. LSTM, with its ability to handle long-range dependencies, and HMM, known for modeling transitions between emotional states, offer complementary strengths. This hybrid model provides a robust framework for understanding and detecting emotions more accurately in real-world environments.

The application of preprocessing techniques like Mel-Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding Coefficients (LPCC) ensures that the most relevant features of speech are extracted, improving the overall system performance. The evaluation of benchmark datasets such as IEMOCAP and RAVDESS emphasizes the hybrid model's superior accuracy compared to traditional methods, particularly in challenging and noisy conditions. However, challenges remain, particularly in addressing issues of bias and fairness in SER systems. Variations in cultural expression, gender, and age can lead to biased outcomes, which require diversified datasets and fairness-aware algorithms to mitigate. Further studies may focus on integrating advanced deep learning techniques like transformer architectures and attention mechanisms to further improve the performance of SER systems. Additionally, multimodal approaches combining audio with visual data hold promise for developing even more accurate and context-aware emotion recognition systems. A total of 1025 Articles were retrieved from various digital library out of which 775 were excluded after the first eligibility criteria, 250 were subjected to the next stage of screening and 200 were discarded due to paywall, or lack of a clear research methodology while others were also discarded due to credible findings of the articles. And subsequently 50 of these articles were subjected for the novel Preferred Reporting Items Metha-Analysis (PRISMA) out of which 30 of the articles were carefully utilized as reported in table 1. due to their clear methodology and potential research gap for further investigation.

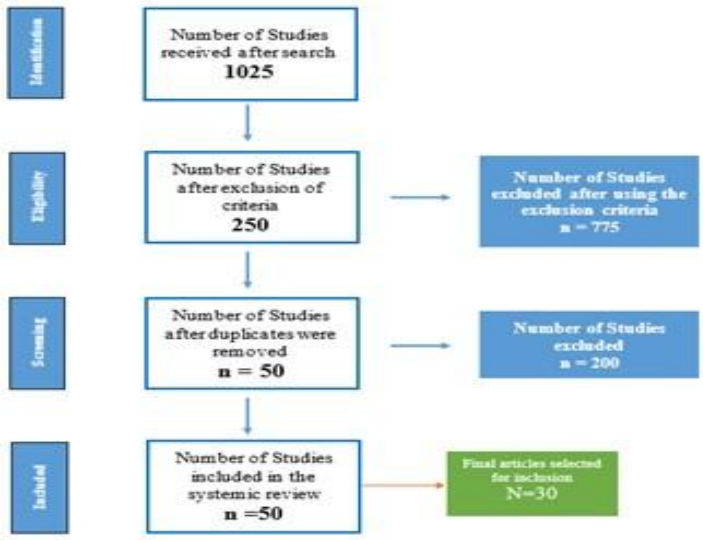

Figure 4: Prisma Analysis Feedback

The number of studies after the exclusion of irrelevant papers were 50, however some of the papers have unclear methodology while others were already old literatures, thus selecting best choice of 30 papers out of the 50 for the research.

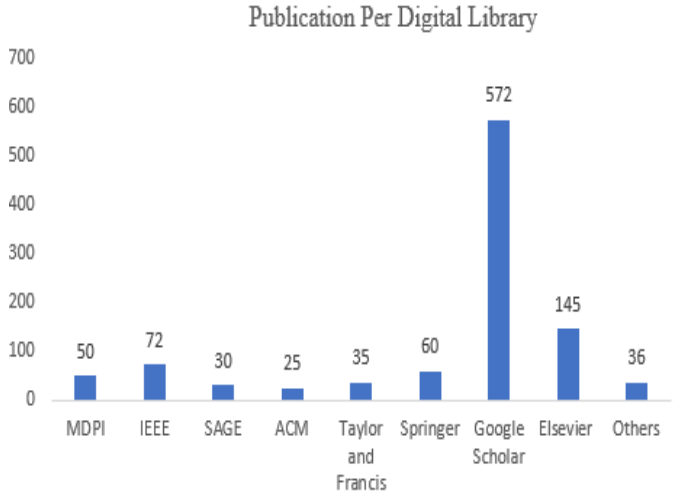

Figure 5: Illustrate Publication Per Digital Library

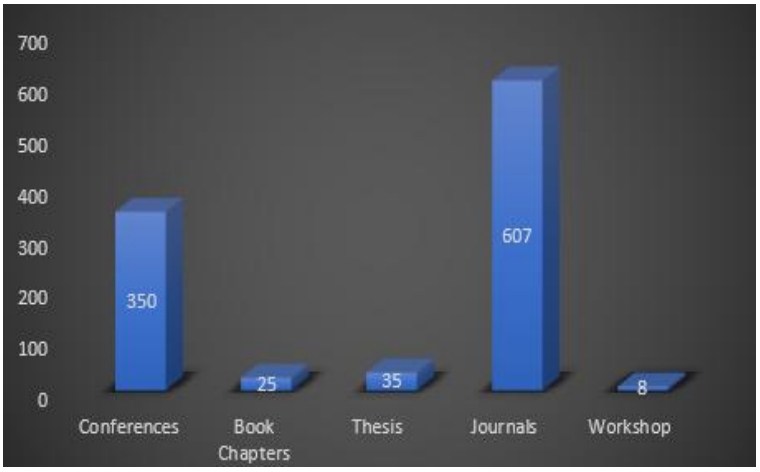

Figure 6: Analysis of Publications from Conferences, Journals, Thesis, Workshops and book Chapters

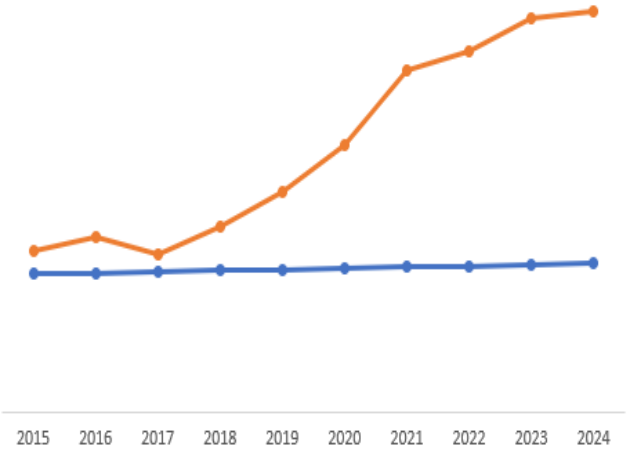

Figure 7: Illustrate Percentage Analysis for the Ten-year Period

The figure 7 depicts the actual states of number of publications and interest of the research community on speech Emotion recognition (SER) within the Ten-year period from 2015 to 2024. It was observed that from 2015 t0 2019 there was an inconsistency in research on SER. However. From 2020 to 2024 there is a drastic shift of researchers from the research community on SER.

CONCLUSION

Speech Emotion Recognition (SER) has an important role in our everyday applications, from mental health diagnostics to human-computer interaction. Thus, Stacked Long Short-Term Memory and Hidden Markov Model for Speech Emotion Recognition-A Systematic Review is developed. The study emphasized the vital place of the integration of Stacked Long Short-Term Memory (LSTM) networks and Hidden Markov Models (HMM) for SER, and addressed the limitations of traditional models. developed hybrid model leverages LSTM's strength in handling sequential data as well as HMM's ability to model emotional transitions, making it highly suitable for real-world noisy environments. Preprocessing techniques such as MFCC and LPCC are applied to enhance feature extraction, and benchmark datasets such as IEMOCAP and RAVDESS are used for evaluation. The research highlights the superior role of the hybrid model's performance on SER and sets the stage for a significant shift in future research in addressing bias and fairness in SER systems by combining LSTM and HMM as hybrid model.

Future Directions in SER Using Advanced Deep Learning and Hybrid Models

The future of Speech Emotion Recognition lies in the continued integration of advanced deep learning techniques, such as transformer-based architectures and attention mechanisms. Transformers, which have revolutionized natural language processing tasks, are now being applied to SER due to their ability to capture long-range dependencies in data more effectively than LSTMs. Attention mechanisms allow the model to focus on the most relevant parts of the speech signal, improving both speed and accuracy. Additionally, there is growing interest in multimodal emotion recognition, where audio data is combined with visual cues like facial expressions or body language to create more accurate emotion recognition systems. Hybrid models that integrate these new deep learning techniques with traditional methods like HMM are expected to offer the best performance in complex, real-world applications. Emerging trends also include the use of unsupervised learning to reduce the reliance on labeled data, making it easier to develop robust models even when high-quality datasets are not available**.**

REFERENCES

Abbaschian, B.J.; Sierra-Sosa, D.; Elmaghraby, A. (2021). Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models. *Sensors*, 21(4), 1249. https://doi.org/10.3390/s21041249

Abbaschian, B.J.; Sierra-Sosa, D.; Elmaghraby, A. (2021). Deep Learning Techniques for Speech EmotionRecognition, from Databases to Models. *Sensors*, 21(4), 1249. https://doi.org/10.3390/s21041249

Ahmed, M. R., Islam, S., Islam, A. K. M. M., & Shatabda, S. (2023). An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition. *IEEE*.

Alhasan, S., Akinyemi, A. E., & Wisdom, D. D. (2020). A comparative performance study of machine learning algorithms for efficient data mining management of intrusion detection systems. *International Journal of Engineering Applied Sciences and Technology*, 5(6), 85-110. ISSN 2455-2143.

Amami, R. (2023). A robust voice pathology detection system based on the combined BiLSTM–CNN architecture. MENDEL Soft Computing Journal, 29 (2), 202–212. https://doi.org/10.13164/mendel.2023.k.202

Bakker, I., Van der Voordt, T., Vink, P., & De Boon, J. (2014). Pleasure, arousal, dominance: Mehrabian and Russell revisited. Current Psychology, 33(3), 405-421.

Chamishka, S., Madhavi, I., Nawaratne, R., Alahakoon, D., De Silva, D., Chilamkurti, N., & Nanayakkara, V. (2022). A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling. *Multimedia Tools and Applications*, 81, 35173–35194. https://doi.org/10.1007/s11042-022-13363-4

Chen, C., & Zhang, P. (2024). TRNet: Two-level refinement network leveraging speech enhancement for noise robust speech emotion recognition. arXiv. https://arxiv.org/abs/2404.12979arXiv

Coto-Jiménez, M. (2021). Discriminative multi-stream post-filters based on deep learning for enhancing statistical parametric speech synthesis, in Biomimetics, MDPI, 6(1), 12. https://doi.org/10.3390/biomimetics6010012

Gendron, B., & Guibon, G. (2024). SEC: Context-aware metric learning for emotion recognition in conversation. Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA 2024), 8–18. https://aclanthology.org/2024.wassa-1.2

Haque, M. M., Islam, S., & Sadat, A. J. M. (2023). Capturing spectral and long-term contextual information for speech emotion recognition using deep learning techniques *(B.Sc. thesis, Islamic University of Technology). Department of Computer Science and Engineering.*

Harby, F., Alohali, M., Thaljaoui, A., & Talaat, A. S. (2024). Exploring Sequential Feature Selection in Deep Bi-LSTM Models for Speech Emotion Recognition. *Computational and Mathematical Methods in Medicine*, 78(2), 2716. https://doi.org/10.32604/cmc.2024.046623

Huang, J., Liu, B., & Tao, J. (2021). Learning long-term temporal contexts using skip RNN for continuous emotion recognition. *Virtual Reality & Intelligent Hardware*, *3*(1), 55-64. https://doi.org/10.1016/j.vrih.2020.11.005.

Jalal, M. A., Loweimi, E., Moore, R. K., & Hain, T. (2019). Learning temporal clusters using capsule routing for speech emotion recognition. *In Proceedings of Interspeech 2019* (pp. 1701-1705). ISCA. https://doi.org/10.21437/interspeech.2019-3068.

Le, D., Aldeneh, Z., & Mower Provost, E. (2017). Discretized continuous speech emotion recognition with multi-task deep recurrent neural network. *Proceedings of the Interspeech Conference*, 481-485. https://doi.org/10.21437/Interspeech.2017-1050.

Lee, J., & Tashev, I. (2015). High-level feature representation using recurrent neural network for speech emotion recognition. *Proceedings of the Interspeech Conference*.

Lim, W., Jang, D., & Lee, T. (2021). Speech emotion recognition using convolutional and recurrent neural networks. *Audio and Acoustics Research Section, ETRI, Daejeon, Korea.*

Lin, W.-C., & Busso, C. (2020). An efficient temporal modeling approach for speech emotion recognition by mapping varied duration sentences into a fixed number of chunks. In *Proceedings of the MSP-Podcast dataset*. Multimodal Signal Processing (MSP) Lab, The University of Texas at Dallas.

Lin, Z., Cruz, F., & Sandoval, E. B. (2024). Self-context-aware model (SCAM) for intelligent interaction. arXiv. https://arxiv.org/abs/2401.10946

Lin, Z., Cruz, F., & Sandoval, E. B. (2024). Self-context-aware model (SCAM) for intelligent interaction. arXiv. https://arxiv.org/abs/2401.10946

Mohammed, M. M., & Schuller, B. W. (2020). ConcealNet: An end-to-end neural network for packet loss concealment in deep speech emotion recognition. arXiv preprint arXiv:2005.07777.

Mu, Y., Hernández Gómez, L. A., Cano Montes, A., Alcaraz Martínez, C., Wang, X., & Gao, H. (2017). Speech emotion recognition using convolutional-recurrent neural networks with attention model. In *Proceedings of the 2017 2nd International Conference on Computer Engineering, Information Science and Internet Technology (CII 2017)* (ISBN: 978-1-60595-504-9).

Nakisa, B., Rastgoo, M. N., Rakotonirainy, A., Maire, F., & Chandran, V. (2018). Long short term memory hyperparameter optimization for a neural network based emotion recognition framework. *IEEE Access*, *6*, 8670-8681. https://doi.org/10.1109/ACCESS.2018.2868361.

Nam, H.-J., & Park, H.-J. (2024). Speech emotion recognition under noisy environments with SNR down to –6 dB using multi-decoder Wave-U-Net. Applied Sciences, 14(12), 5227. https://doi.org/10.3390/app14125227

Patel, D., Amipara, S., Sanaria, M., Pareek, P., Jayaswal, R., & Patil, S. (2024). ASER: An exhaustive survey for speech recognition based on methods, datasets, challenges, and future scope. https://doi.org/10.18280/ria.380218.

Shahin, I., Hindawi, N., Bou Nassif, A., Alhudhaif, & Polat, K. (2023). Novel dual-channel long short-term memory compressed capsule networks for emotion recognition.

Sönmez, Y. Ü., & Varol, A. (2020). A speech emotion recognition model based on multi-level local binary and local ternary patterns. *IEEE Access*, 8, 187110-187121. https://doi.org/10.1109/ACCESS.2020.3031763.

Tanoko, Y., & Zahra, A. (2022). Multi-feature stacking order impact on speech emotion recognition performance. *Bulletin of Electrical Engineering and Informatics*, *11*(6), 3272-3278. https://doi.org/10.11591/eei.v11i6.4287

Tzinis, E., & Potamianos, A. (2017). Segment-based speech emotion recognition using recurrent neural networks. In Proceedings of the IEEE Conference (pp. 1-8).2017 DOI: 10.1109/ACII.2017.8273599

Ullah, R., Asif, M., Shah, W. A., Anjam, F., Ullah, I., Khurshaid, T., Wuttisittikulkij, L., Shah, S., Ali, S. M., & Alibakhshikenari, M. (2023). Speech emotion recognition using convolution neural networks and multi-head convolutional transformers. *Sensors*, *23*(13), 6212. https://doi.org/10.3390/s23136212.

Van Houdt, G., Mosquera, C., & Nápoles, G. (2020). A review on the long short-term memory model. *Artificial Intelligence Review*, *53*(8), 5929-5955. https://doi.org/10.1007/s10462-020-09838-1.

Venkataramanan, K., & Rajamohan, H. R. (2019). Emotion recognition from speech. arXiv preprint arXiv:1912.10458.

Wang, L., Zhang, S., & Liu, Y. (2025). Graph attention model with contextual reasoning and emotion-shift awareness for conversational emotion recognition. Complex & Intelligent Systems, 1–14. https://doi.org/10.1007/s40747-025-01903-y

Wu, X., Cao, Y., Lu, H., Liu, S., Wang, D., Wu, Z., Liu, X., & Meng, H. (2021). Speech emotion recognition using sequential capsule networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 325–336. https://doi.org/10.1109/TASLP.2020.3039432.

Wu, X., Liu, S., Cao, Y., Li, X., Yu, J., Dai, D., Ma, X., Hu, S., Wu, Z., Liu, X., & Meng, H. (2019). Speech emotion recognition using capsule networks. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 5386-4658. https://doi.org/10.1109/ICASSP.2019.8683652.

Zhang, H., Huang, H., & Han, H. (2021). A novel heterogeneous parallel convolution Bi-LSTM for speech emotion recognition. *Applied Sciences*, 11(21), 9897. https://doi.org/10.3390/app11219897

Zhang, H., Li, M., Chen, Y., & Wang, Q. (2024). CLEF: Counterfactual learning framework for debiasing context-aware emotion recognition. Emergent Mind, 1–12. https://www.emergentmind.com/papers/2403.05963

Zhang, C., Yu, J., & Chen, Z. (2021). Music Emotion Recognition Based on Combination of Multiple Features and Neural Network. *2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*. 1-6

Zhao, J., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control*, *47*, 312–323. https://doi.org/10.1016/j.bspc.2018.08.018.

Zhao, L., Xuan, J., Lou, J., Yu, Y., & Yang, W. (2025). Context-aware academic emotion dataset and benchmark (RAER). Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

Zheng, C., Wang, C., & Jia, N. (2019). An ensemble model for multi-level speech emotion recognition. *Information*, 10(12), 394. https://doi.org/10.3390/info10120394