Academic Science Journal (ASJ)
Vol. 4, No. 3, July 2026, pp. 64~77
DOI: 10.24237/04.02.626

An Evaluation Study for Worthwhile Research in Speech Emotion Recognition Using Deep Learning

Samah Abbas Ali, Jamal Mustafa Al-Tuwaijari
Department of Computer Science, College of Science, University of Diyala, Iraq.

Article Info

Article history:

Received 14, 01, 2025
Revised 20, 03, 2025
Accepted 13, 05, 2025
Published 30, 07, 2026

Keywords:

Deep Learning,
SER,
CNN,
LSTM,
RNN.

ABSTRACT

Speech Emotion Recognition (SER) is an emerging field in Human-Computer Interaction (HCI) focused on detecting emotions from speech. This study reviews recent advancements in SER, emphasizing AI methods, particularly deep learning techniques that significantly enhanced emotion detection accuracy. Sources include Google Scholar, IEEE, ResearchGate, and MDPI, providing an in-depth assessment of evolving techniques. Developments in deep learning, especially Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks, improved precision and robustness. Feature extraction techniques such as MFCC and spectrogram analysis, alongside training methods like data augmentation and transfer learning with datasets like RAVDESS and IEMOCAP, are explored. Additionally, hybrid and multimodal approaches combining speech with text and facial expressions are highlighted to enhance recognition accuracy. The paper compares AI models' performance, identifying strengths and areas needing development. This study provides comprehensive insight into current SER progress, summarizes recent techniques, and suggests future research directions to further improve the accuracy and adaptability of SER systems.

Corresponding Author:

Samah Abbas Ali
Department of Computer Science, College of Education for Pure Science,
University of Diyala,
Baqubah City, Diyala Governorate, Iraq.
Email: scicompms232404@uodiyala.edu.iq

1. INTRODUCTION

Speech Emotion Recognition (SER) is an emerging field in Human-Computer Interaction (HCI) that improves interaction by detecting emotions through speech signals. Emotions like happiness, sadness, anger, and surprise are identified using speech features such as melody, rhythm, intonation, and modulation. Deep learning algorithms, especially multilayer perceptrons, play a key role in recognizing emotional patterns in speech data. Recent advancements using CNNs, RNNs, and LSTMs have significantly improved emotion detection accuracy and reliability [1].

SER's ability to interpret emotions is essential for applications in smart voice-activated assistants, gaming, e-learning, and mental health tools. Early SER models relied on basic features like pitch and tone but struggled with variations in individual speech and external factors. The development of algorithms like Support Vector Machines (SVMs) and neural networks improved precision, while CNNs and LSTMs enabled extraction of subtle emotional signals, enhancing accuracy [2].

Currently, SER is applied in healthcare, education, and customer service to improve human interaction with technology. As systems evolve, incorporating multimodal data, including facial expressions and body language, will further improve emotion recognition, leading to more intuitive human-machine communication [2].

Advancements in deep learning techniques, particularly CNNs, RNNs, and LSTMs, continue to enhance SER systems, expanding their use beyond entertainment to critical sectors like healthcare and

education. Looking to the future, combining vocal signals with facial expressions and body language will enhance the accuracy and reliability of emotion recognition, creating richer user experiences [2].

In the past few years, substantial advances have been made in Speech Emotion Recognition (SER) through deep learning strategies. Many researchers have concentrated on creating novel methods and refining existing systems to boost the accuracy and reliability of emotion recognition from speech signals Chen Jie (2021) [2] showcased a CNN-based model that employs MFCC, tested across various datasets, underscoring the need for further stability enhancements. Chintalapudi et al. (2023) [1] implemented CNNs to identify emotions within the Kaggle Speech Recognition dataset, obtaining an F1-score between 60% and 73%, despite facing challenges regarding emotional variability.

2. METHOD

Explaining research chronological, including research design, research procedure (in the form of algorithms, Pseudocode or other), how to test and data acquisition.

This paper reviews developments in Speech Emotion Recognition (SER), with an emphasis on deep learning approaches. The methodology is organized into three distinct stages, as illustrated in Figure 1. This framework offers a thorough insight into the evolution of deep learning methods in SER, ranging from data gathering to concluding analysis and comparison.

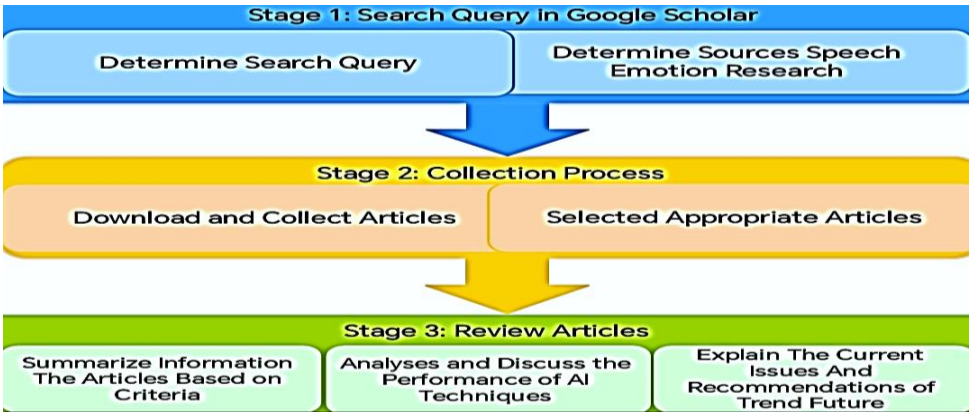

Figure 1. Methodology for Systematic Process

Stage 1: Keyword-Based Search: The first stage involves searching for relevant papers in academic databases like Google Scholar, IEEE Xplore, Research Gate, and MDPI using keywords like "Speech Emotion Recognition","SRE","Deep Learning","CNN in SER", "Hybrid Models in Emotion Detection".

Stage 2: Paper Selection and Filtering: In this stage, papers published between 2019 and 2024 are selected, focusing on those that apply deep learning models in SER, their methodologies, and outcomes.

Stage 3: Review and Analysis: This final stage involves evaluating the selected papers' methodologies, including deep learning models (CNNs, LSTMs, RNNs), feature extraction, training strategies, and datasets to assess accuracy, robustness, and future research opportunities.

Based on findings of studies in Table 1, there hasn't been an all-encompassing analysis of performance across different AI models for Speech Emotion Recognition (SER). This study evaluates accuracy of various artificial intelligence models to ascertain which excel in emotion prediction. The paper offers a comprehensive assessment of several AI models, focusing on commonly employed ones such as artificial neural networks (ANNs), deep neural networks (DNNs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs). Results can guide researchers and practitioners in choosing appropriate models for SER.

Table 1. Studies on Speech Emotion Recognition Using Deep Learning Techniques.

Ref	Methodology	Year	Al Techniques
[3]	CNN and LSTM For SER	2019	CNN, LSTM
[4]	End-to-End SER With DNNs	2019	CNN, LSTM
[2]	CNN For SER With MFCC	2021	CNN, MFCC
[5]	Robust SER With CNN, LSTM	2021	CNN+LSTM, SFS
[1]	CNN For SER	2023	CNN

2.1. Speech Emotion Recognition and Al Approaches

Speech Emotion Recognition (SER) has recently attracted considerable interest, largely because deep learning methods can significantly improve its accuracy and reliability. The deep learning models most frequently utilized in SER include Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and Recurrent Neural Networks (RNNs). These models utilize deep learning's capacity to recognize intricate patterns and relationships in speech data, making them especially effective for tasks involving emotion classification.

2.2. Convolutional Neural Networks (CNNs)

CNNs have become widely used in SER for extracting features from spectrograms or direct speech signals. CNNs excel at identifying localized patterns in speech, such as variations in intensity and pitch, which are essential for recognizing different emotional states. Recent research has demonstrated that CNNs surpass traditional machine learning approaches due to their capability to automatically derive pertinent features from raw audio data without the need for manual feature extraction. Furthermore, CNNs are highly proficient in managing high-dimensional data, making them an ideal option for real-time SER applications. The accuracy of CNN-based models in SER ranges from 46.85% to 99.47% [6][7][8][9][10][11][12][13][14][15][16][17][18][19][20][21][22].

2.3. Recurrent Neural Networks (RNNs)

RNNs, especially variants like GRUs (Gated Recurrent Units), are created to handle sequential data by preserving the memory of prior inputs. In speech emotion recognition (SER), RNNs are employed to model the dynamic characteristics of speech signals. By analyzing speech sequences frame by frame, RNNs can monitor emotional changes and identify subtle tone variations that reflect shifts in emotion. Architectures based on RNNs are especially effective in managing longer sequences, enhancing the model's capability to detect emotions over time. The accuracy levels of RNN-based models in SER range from 72% to 97% [13][21][23][24].

2.4. Long Short-Term Memory (LSTM) Networks

LSTM networks, a variant of Recurrent Neural Networks (RNN), are designed to capture long-term dependencies in data. In Speech Emotion Recognition (SER), LSTMs excel at handling temporal aspects of speech, such as rhythm, intonation, and prosody, crucial for emotion detection. LSTM-based models can distinguish similar emotions using temporal indicators. Recent developments have integrated LSTMs with Convolutional Neural Networks (CNNs) to improve performance by merging spatial and temporal features. Accuracy of LSTM-based models in SER ranges from 74.96% to 99.5% [7][18][19][23][24][25][26][27].

2.5. Hybrid Approaches

An increasingly popular strategy for Speech Emotion Recognition (SER) is a hybrid model that merges CNNs, LSTMs, and RNNs. This combination utilizes the unique strengths of each architecture: CNNs excel in feature extraction, LSTMs capture temporal dependencies, and RNNs are effective for sequential modeling. Impressive results have been demonstrated by these hybrid models in terms of classification accuracy and resilience, particularly when used with extensive and varied datasets. The accuracy of hybrid-based models in SER ranges from 85.39% to 100% [3][10][28][29][30].

2.6. Machine Learning Techniques for Speech Emotion Recognition

Alongside deep learning algorithms, traditional machine learning methods have been extensively utilized in speech-emotion recognition systems have been featured [20][27][31], showing accuracy rates between 65.43% and 95.1%.

2.7. Feature Extraction Methods in SER

The process of feature extraction is essential for the effectiveness of SER models. Efficient feature extraction techniques assist in converting raw speech signals into a collection of meaningful features that machine learning algorithms can utilize for classification.

2.8. Mel-Frequency Cepstral Coefficients (MFCCs)

MFCCs are widely used features in speech emotion recognition (SER), capturing the speech signal's power spectrum and aligning closely with human speech perception. They help SER models focus on important acoustic features, such as formants, pitch, and intensity, essential for emotion classification. MFCCs effectively identify a wide range of emotions and are often combined with deep learning models like CNNs and LSTMs [32][33][34].

2.8.1. Spectrogram Analysis

Spectrograms illustrate the frequency characteristics of speech over time and are commonly used as input features for deep learning models. CNNs can extract spatial features from spectrograms that reflect changes in speech signals across time. The visual representation helps models identify patterns related to emotional states, such as vocal fry or pitch variations, indicating emotions like anger or joy [6][35].

2.8.2. Raw Speech Signals

Raw speech signals can be used as inputs for deep learning models in end-to-end architectures without relying on hand-crafted features. This approach allows models to automatically identify essential features during training, making it increasingly popular. By analyzing raw speech, models can capture low-level characteristics that traditional feature extraction methods might miss, enhancing the accuracy of emotion classification [36][37][38].

2.9. Training Techniques for SER Models

Training methods are crucial for optimal performance in Speech Emotion Recognition (SER). Techniques like supervised and unsupervised learning, data enhancement, and transfer learning are frequently employed to boost model generalization and minimize overfitting.

2.9.1. Supervised Learning

Supervised learning is the primary method used in SER, utilizing labeled datasets to train models linking audio samples to specific emotion labels. This method is effective with datasets like TESS and IEMOCAP and significantly improves SER accuracy, especially when combined with advanced deep learning techniques like CNNs and LSTMs [39][40].

2.9.2. Unsupervised Learning

Unsupervised learning methods, such as clustering and dimensionality reduction, have been explored for SER. These techniques allow models to detect patterns without labeled information, making them valuable when labeled data is scarce or expensive. Recent studies suggest combining unsupervised and supervised learning improves the resilience and flexibility of SER systems [41][42].

2.10. Data Augmentation

Data augmentation expands datasets by applying transformations to existing data. In SER, this includes adding noise, changing pitch or speed, and varying vocal tract length. These diverse examples help models generalize better and reduce overfitting, particularly when training deep learning models on smaller datasets [43][44].

2.11. Datasets Used in SER

Emotional datasets are crucial in SER research, aiding model training and assessment across scenarios. They vary in size and type, including speech, facial expressions, and gestures, helping develop models for complex situations. Dataset details are in Table 2.

Table 2. Data set in SER

Data Set Name [Ref]	Language	Data Count	Data Type	Number of Subjects	Number of Emotions	Emotions
RML– Ryerson Multimedia Laboratory [45]	English	728 samples	Audio-Visual	24 speakers	8	Happiness, Sadness, Anger, Fear, Surprise, Disgust, Calm, Neutral
Emo-DB – Berlin Database of Emotional Speech [46]	German	535 samples	Audio	10 speakers	7	Anger, Disgust, Fear, Happiness, Sadness, Surprise, Neutral
SAVEE – Surrey Audio-Visual Expressed Emotion Database [47]	English	480 samples	Audio-Visual	4 speakers	7	Anger, Disgust, Fear, Happiness, Sadness, Surprise, Neutral
IEMOCAP – Interactive Emotional Dyadic Motion Capture [48]	English	12 Hours approx. (12,000) samples	Audio-Visual	10 speakers	5	Anger, Happiness, Sadness, Neutral, Frustration
CMU-MOSEI – Carnegie Mellon University Multimodal Opinion Sentiment and Emotion Intensity [49]	English	23,500 samples	Audio-Visual	More Than 1000 speakers	7	Happy, Angry, Sad, Surprise, Disgust, Fear, Neutral
RAVDESS – Ryerson Audio-Visual Database of Emotional Speech and Song [50]	English	1,440 samples	Audio-Visual	24 speakers	8	Anger, Disgust, Fear, Happiness, Sadness, Surprise, Calm, Neutral
TESS – Toronto Emotional Speech Set [51]	English	2,800 samples	Audio	2 speakers	7	Anger, Disgust, Fear, Happiness, Sadness, Surprise, Neutral
CREMA-D – Crowd-sourced Emotional Multimodal Actors Dataset [52]	English	7,442 samples	Audio-Visual	91 speakers	6	Anger, Disgust, Fear, Happiness, Sadness, Neutral
MSP-Podcast – Multimodal Sentiment Analysis in Real-life Media [53]	English	237 Hours samples	Audio-Visual	151,654 speakers	7	Anger, Happiness, Sadness, Neutral, Surprise, Disgust, Fear
FAU-Aibo – Emotion in the Speech of Children [54]	German	2,500 samples	Audio	51 children	5	Anger, Happiness, Sadness, Neutral, Fear
BAVED – Baghdad Audio-Visual Emotional Dataset [55]	Arabic	3,000 samples	Audio-Visual	20 speakers	6	Anger, Happiness, Sadness, Neutral, Disgust, Fear

Korean Speech Emotion Database (Chosun University) [56]	Korean	2,880 recordings	Audio	Multiple speakers	4	Anger, Happiness, Sadness, Neutral
PROMPT—Prosodic and Emotion-Oriented Speech Corpus. [57]	English	160,124 Samples	Audio, Text	Multiple speakers	7	Anger, Disgust, Fear, Happiness, Sadness, Surprise, Neutral

2.12. Evaluation Metrics for SER Models

The effectiveness of Speech Emotion Recognition models is evaluated using metrics like accuracy, robustness, and real-world relevance, assessing their ability to identify emotions and perform well under varying conditions.

2.12.1. Accuracy

Accuracy assesses the ability of the model to correctly classify emotions. It is the most frequently utilized evaluation metric and is crucial for evaluating the performance of SER models. Elevated accuracy suggests that the model can reliably differentiate between emotions like happiness, anger, and sadness [58].

2.12.2. Robustness

Robustness refers to a model's ability to maintain strong performance despite variations in speech data, such as background noise or different speaker traits. This quality is crucial for real-world applications, where speech conditions can vary widely. Improving robustness is a key area of research in Speech Emotion Recognition (SER), often using noise-resistant feature extraction and hybrid deep learning techniques [59][60].

2.12.3. Real-World Applicability

The assessment of real-world applicability determines how effectively a model functions in practical situations. This includes examining real-time processing, adaptation to languages or accents, and integration with systems like virtual assistants or customer service platforms. Adaptability is vital for implementing SER models in commercial contexts [26][45][60].

3. RESULTS AND DISCUSSION

This section reviews the results in Speech Emotion Recognition (SER) system, methods applied, and accuracy levels from 47 scholarly articles published between 2019 and 2024 Tables 3, 4, 5, 6, and 7. The studies utilized various data types, including audio, video, image, text, and multimodal strategies. The distribution of data types is visualized in the chart below (Figure 2): Audio was the most common data type, present in 70% of the studies, followed by video (15%), image (10%), text (5%), and multimodal methods (25%).

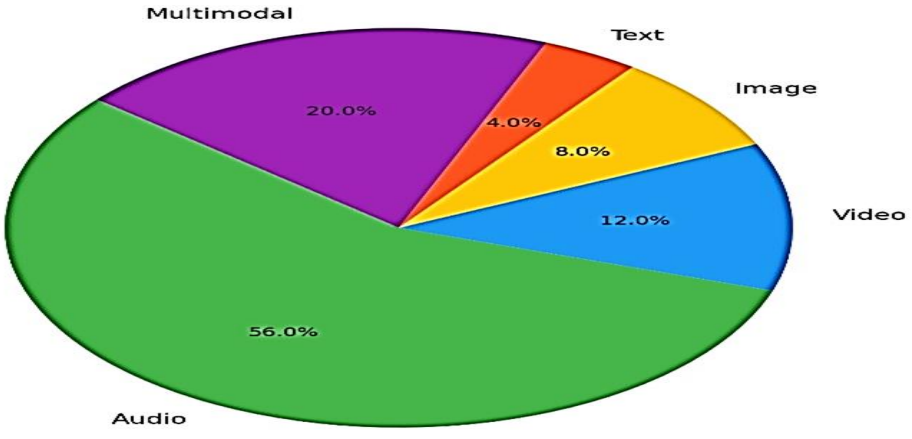

Figure 2. Data Types Distribution in SER Studies

The studies reviewed employed various deep learning and machine learning methods, as shown in the chart (Figure 4): deep learning was used in 80% of studies, machine learning in 15%, and other techniques in 5%. Deep learning methods are considered superior due to their ability to automatically extract hierarchical features and model complex patterns in raw speech data, which leads to higher accuracy and better generalization across diverse datasets.

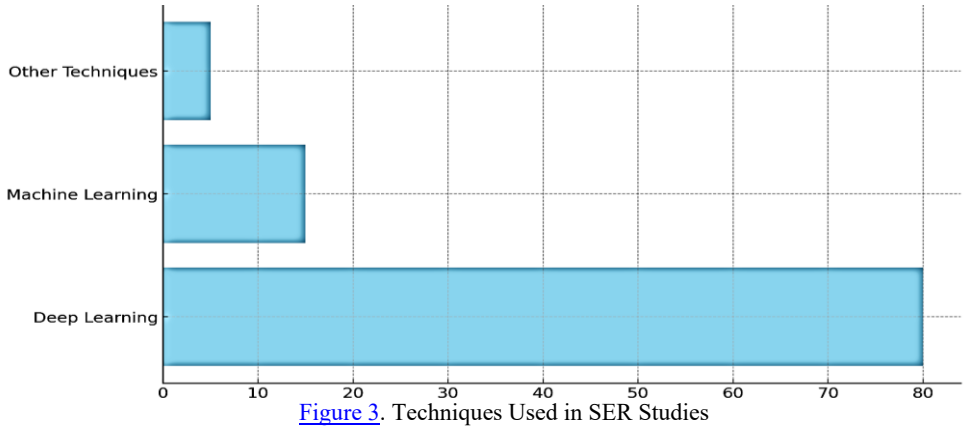

Figure 3. Techniques Used in SER Studies

The accuracy of SER models varies with feature extraction techniques. Figure 4 shows that MFCC was used in 40% of studies, spectrogram analysis in 30%, raw speech signals in 20%, and other techniques in 10%. MFCC is preferred because it effectively simulates the human auditory system, capturing the most relevant frequency components for emotional cues in speech. It provides a compact, discriminative representation that performs consistently well across different SER tasks.

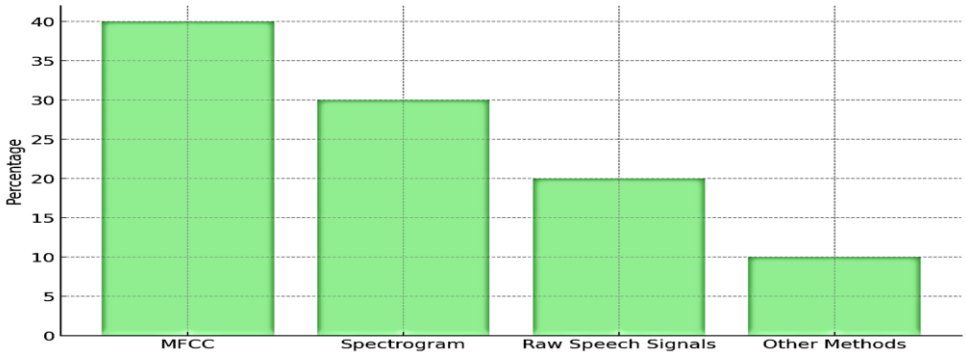

Figure 4. Accuracy Distribution Based on Feature Extraction Methods

The chart in Figure 5 illustrates accuracy distribution by technique: 30% of studies had high accuracy (90–100%), 50% had moderate accuracy (70–89%), and 20% had low accuracy (50–69%). Among evaluation metrics, accuracy is the most commonly used because it directly reflects how well a model distinguishes between different emotional classes. It is easy to interpret and provides a clear benchmark for comparison between models.

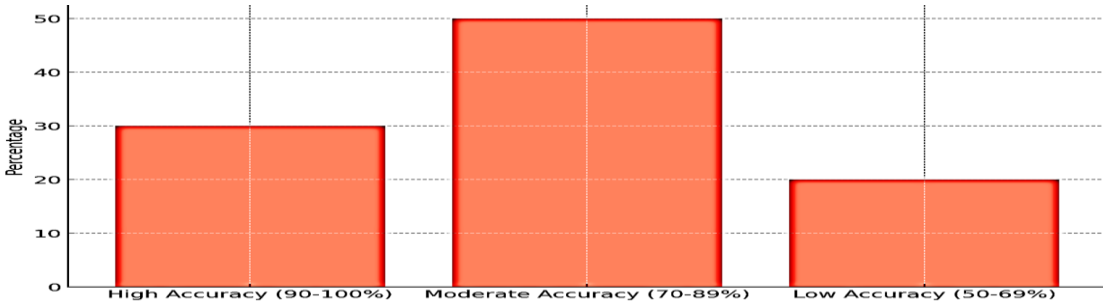

Figure 5. General Accuracy Distribution of SER Techniques

Deep learning models, such as CNNs and LSTMs, significantly outperform traditional machine learning algorithms in accuracy. Hybrid methods using various neural networks also demonstrate strong performance. Figures 6 and 7 illustrate the usage percentage and accuracy ranges of these algorithms. This performance advantage is largely due to the deep models' ability to capture both spatial and temporal dynamics of speech signals, enabling more nuanced emotional classification.

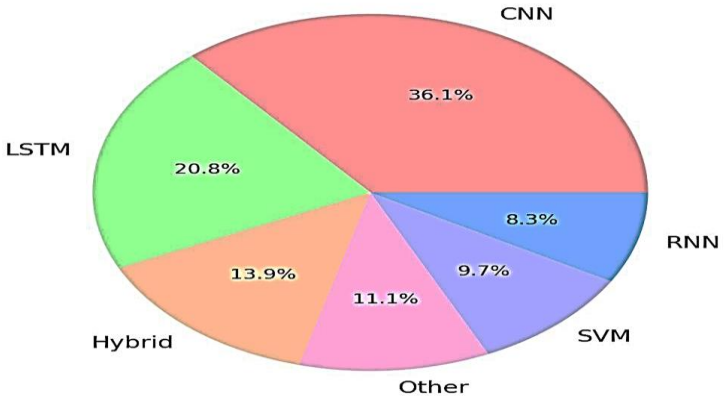

Figure 6. Usage Percentage of Algorithms in SER

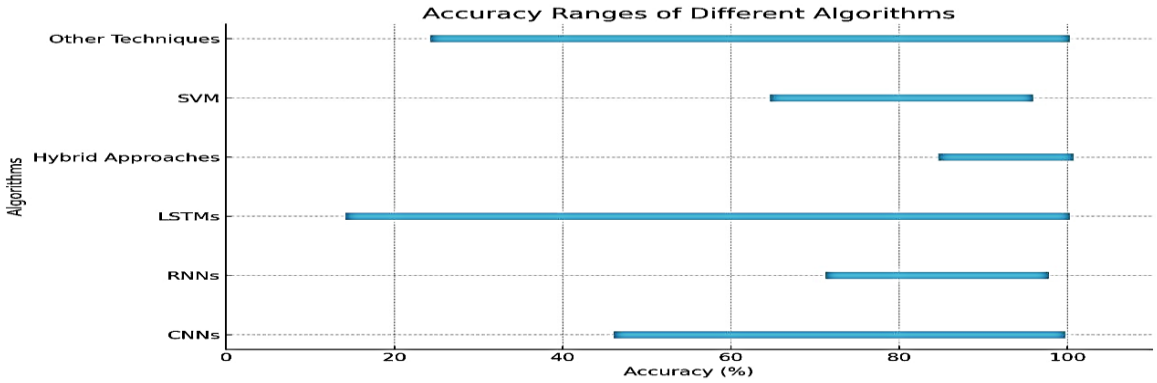

Figure 7. Accuracy Ranges of Different Algorithms in SER

In terms of training techniques, supervised learning remains the preferred approach, as it provides direct mapping between labeled speech samples and their corresponding emotional states. This leads to high performance when large labeled datasets like RAVDESS and IEMOCAP are available. Although unsupervised learning methods are valuable when labeled data is limited, they typically underperform compared to supervised models in classification tasks.

Table 3. Methods of Different Al-techniques For SER in 2019.

Ref	Year	Method	Feature Extraction	Dataset	Accuracy
[6]	2019	Deep learning: ADRNN (Attention-based Dilated Residual Neural Network)	Attention Mechanism, Dilated CNN	IEMOCAP, EMO-DB	90.78% (speaker-dependent) and 85.39% (speaker-independent) for EMODB,74.96% unweighted accuracy (speaker-dependent) 69.32% unweighted accuracy (speaker-independent) for IEMOCAP
[23]	2019	Deep learning: CNN, LSTM, RNN	MFCC, noise reduction	(Emo-DB), (IEMOCAP)	High accuracy (80-90%)

[28]	2019	Deep learning: 1D & 2D CNN LSTM	Iog-Mel Spectrograms, LFLB	Berlin EmoDB, IEMOCAP	Highest resolution 95.89% (EmoDB, speaker-independent)

Table 4. Methods of Different Al-techniques For SER in 2020.

Ref	Year	Method	Feature Extraction	Dataset	Accuracy
[7]	2020	Deep Learning: (CNN), (BiLSTM).	Spectrogram Conversion Using (STFT)	IEMOCAP, EMO-DB, RAVDESS	EMO-DB: 85.57% RAVDESS: 77.02% IEMOCAP: 72.25%
[8]	2020	Deep Learning:CNN	LIBROSA	RAVDESS	82%
[9]	2020	Deep Learning: (CNN), (DSCNN)	Spectrograms	SAVEE	DSCNN: 87.8% CNN: 79.4%
[10]	2020	Deep Learning: The model combines Local Feature Learning Blocks (LFLBs) with parallel CNNs of varying filter lengths, followed by LSTM for global contextual information.	LFLB, Parallel CNNs	(EmoDB) and RAVDESS dataset.	The model achieved state-of-the-art accuracy with three parallel CNNs and modified LFLBs using average pooling and ReLU activation.
[31]	2020	Deep Learning: SVM	MFCC, ZCR HNR, TEO, AE	RML	65.43% (without AE), 72.83% with basic AE, 74.07% (with stacked AE)
[32]	2020	Speech Emotion Recognition: SVM	MFCC, LPCC	LDC, UGA database	"85.08% overall LDC dataset: 90.08% UGA dataset: 65.95%
[33]	2020	Deep Learning,Various methodologies for speech emotion recognition (HMM) (GMM),(SVM), (ANN)	MFCC, LPCs, HNR, TEO	acted, elicited, and natural speech databases.	80%
[61]	2020	Various (Speech, Facial, Text): Stationary Wavelet Transform (SWT)،Particle Swarm Optimization (PSO) assisted Biogeography-based optimization (BBO)	SWT, PSO-BBO	JAFFE, CK+, Berlin EmoDB, SAVEE	99.47% (PSO-BBO for speech recognition) ،Stationary Wavelet Transform for facial recognition: 98.83%،Statistical features with physiological signals: 87.15% Rough set theory with SVM for text semantics: 87.02%

Table 5. Methods of Different Al-techniques For SER in 2021.

	Year	Method	Feature Extraction	Dataset	Accuracy
[11]	2021	Deep Learning: (CNN), (LSTM) Networks, Multi-Layer Perceptron (MLP)	MFCC Used Librosa	RAVDESS, SAVEE	CNN: 47%, MLP: 25%, LSTM: 15%
[12]	2021	Deep Learning: Including CNN, LSTM, GAN, and autoencoders	MFCC	Various datasets including EMO-DB, IEMOCAP, RAVDESS, TESS, CREMA-D	Varies across different datasets and methods; specific accuracies are provided for different methods within the review. For example: EMO-DB: up to 96.97%

[24]	2021	Deep Learning: (DCNN), (RNN), LSTM	MFCCs, TEO	IEMOCAP, CMU-MOSEI, CMU-MOSI, Berlin Emotional	80.51% (DCCN) with SincNet layer, 73.98% for LSTM MMAN
[25]	2021	Deep Learning: 2D Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM)	MFCC	SAVEE, RAVDESS, TESS, CREMA-D,along with an unseen dataset, EMO-DB.	validation accuracy of 67.58%, testing accuracy of 71.28%
[29]	2021	Deep Learning: CNN-RNN Hybrid Architecture	MFCC	(RAVDESS)	98.23%
[34]	2021	Traditional:Support Vector Machines (SVM), Deep Learning: Deep Neural Networks (DNN)	MFCCs,TEO, HNR,Prosodic Features	Berlin Emo-DB, Chinese databases, SAVEE, etc.	SVM up to 95.1%, DNN 97.1%
[47]	2021	Transfer Learning: Dense Neural Networks	Wav2vec 2.0	IEMOCAP, RAVDESS	67.2% (IEMOCAP), 84.3% (RAVDESS)

Table 6. Methods of Different Al-techniques For SER in 2023.

Ref	Year	Method	Feature Extraction	Dataset	Accuracy
[13]	2023	Representation learning,deep learning: (DNNs), (CNNs), (RNNs), autoencoders (AEs), variational autoencoders (VAEs), generative adversarial networks (GANs)	DNN, CNN, RNN, AE, VAN, GAN	IEMOCAP, MSP-IMPROV, EMO-DB, etc.	Accuracy Varied depending on the dataset and method and from it High accuracy 86.5%
[14]	2023	Machine Learning:Random Forest (RF), Deep Learning:Convolutional Neural Networks (CNN)	MFCC	RAVDESS, CREMA-D, TESS, and SAVEE	Exact accuracy values aren't specified, but the comparison between CNN and RF models is discussed.
[15]	2023	Deep Learning,Multi-Head Convolutional Transformer: (CNNs), Transformer Encoder	AWGN, CNN	RAVDESS and IEMOCAP	82.31% on the RAVDESS dataset for eight emotions 79.42% on the IEMOCAP dataset for five emotions
[16]	2023	speech emotion recognition: (CNN)	MFCC,Gender-Dependent Approach	(RAVDESS)	average accuracy of 72.07% for 8 emotions
[17]	2023	Deep Learning, Speech Processing: (CNN)	AlexNet, VGG16	RAVDESS and SAVEE	AlexNet: RAVDESS 87.31%, SAVEE 85.37% VGG16: RAVDESS 85.72%, SAVEE 82.92%
[26]	2023	Deep Learning: CNN + BiLSTM	MFCC, ZCR, RMS	TESS, EmoDB, RAVDESS	100% (TESS), 99.5% (EmoDB), 90.12% (RAVDESS)
[27]	2023	Machine Learning: LSTM network, SVM	HOG, UA, Face Mesh	PROMPT	HOG features: 79% AU features: 71% Face mesh features: 66%
[30]	2023	Deep learning:	MFCC, GTCC, ERB spectrogram.	Korean speech emotion	96% for the two-stream model (Bi-LSTM + YAMNet)

		Bi-LSTM and CNN based transfer learning model			
[62]	2023	Deep Learning: Tensor Fusion Network (TFN), Low-rank Multimodal Fusion (LMF), Mutual Information Maximization (MMIM), among others.	(Speech، Text، Facial، Fusion) Features	IEMOCAP, YouTube, MOUD, ICT-MMMO, CMU-MOSI, CMU-MOSEI, OMG, MELD, SEWA, CH-SIMS	Not specified (since it is a survey paper, it discusses accuracy for different methods but doesn't provide a single accuracy value).

Table 7. Methods of Different Al-techniques For SER in 2024.

Ref	Year	Method	Feature Extraction	Dataset	Accuracy
[18]	2024	Deep Learning, Transfer Learning:(CNN), Spatial Transformer Networks, (BiLSTM)	CNN, LSTM	RAVDESS dataset	The goal is to achieve improved accuracy and robustness compared to traditional single-modality models
[19]	2024	Deep Learning: (CNN), (LSTM)	MFCC	EMO-DB, CASIA	85.72% (EMO-DB), 82.86% (CASIA)
[20]	2024	Machine Learning: (SVM), Decision Tree, (MLP) Deep Learning: (CNN), (LSTM)	MFCC	RAVDESS, TESS	87% (CNN), 85.71% (MLP),81.56% (SVM), 82%(LSTM),72% Decision Tree
[21]	2024	Deep Learning: Various methods including CNN, RNN, LSTM, DBN	CNN, RNN, LSTM, DBN.	EMODB, SAVEE, RAVDESS	Varies based on the dataset and model used (specific values not provided)
[22]	2024	Deep Learning: (CNN)	MFCC, Mel spectrogram, RMS, ZCR.	BAVED dataset	Emotion Recognition: 93% Gender Recognition: 99%
[38]	2024	Deep Learning: DCNN, CapsNet, AlexNet, VGGNet, ResNet, DenseNet, GoogLeNet, and EfficientNet	Involves Transforming Raw data in to useful characteristics for a model.	RAVDESS, IEMOCAP, EMO-DB, TESS, SAVEE, CREMA-D, AESDD, eNTERFACE, EMOVO, MELD, DES, MSP-Podcast, and FAU-Aibo	Accuracy Varied by model and dataset from it WMFCCVQ-RBFN on EMO-DB, SAVEE: 93.6%, 91.82%,1D, 2D CNN-LSTM on EMO-DB, IEMOCAP: 95.33%, 89.16%
[48]	2024	Deep Learning, Transfer Learning "Transfer Learning Models "	Structural Features، Explainable AI	RAVDESS, CREMA-D, TESS, and SAVEE	87%

4. CONCLUSION

This study highlights advancements in Speech Emotion Recognition (SER) using deep learning from 2019 to 2024. A review shows that CNNs, LSTMs, and RNNs have significantly improved SER accuracy, approaching human-level performance. Techniques like transfer learning, data augmentation, and hybrid methods have addressed issues of data variability and noise.

However, challenges remain, such as handling dialect differences, reducing reliance on annotated datasets, and improving model interpretability. Future research should focus on integrating multimodal data (audio, images, text) and developing systems that adapt to changing conditions. This research serves as a

valuable resource for scholars and practitioners, outlining trends in SER and highlighting areas for further study.

ACKNOWLEDGEMENTS

The authors would like to formally acknowledge and express their sincere appreciation to all those who contributed to the successful completion of this paper. Special thanks are extended to colleagues, reviewers, and affiliated institutions for their valuable support, guidance, and constructive feedback throughout the preparation of this work.

REFERENCES

[1] K. S. Chintalapudi, V. S. Muvvala, I. A. Khan Patan, S. V. Gangashetty, H. V. Sontineni, and A. K. Dubey, "Speech Emotion Recognition Using Deep Learning", 2023, International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, Jan. 2023, pp. 1–8, DOI: 10.1109/ICCCI56745.2023.10128612.

[2] J. Chen, "Speech Emotion Recognition Based on Convolutional Neural Network", 2021, International Conference on Networking, Communications and Information Technology (NetCIT), pp. 106-108, 2021, DOI: 10.1109/NetCIT54147.2021.00028.

[3] C. H. Chen, C. Y. Chen, and C.-H. Chen, "Speech Emotion Recognition Using CNN and LSTM", IEEE Transactions on Affective Computing, 2019.

[4] S. R., S. M., and S. M., "End-to-End Speech Emotion Recognition Using Deep Neural Networks", Int. Res. J. Modernization Eng. Technol. Sci. (IRJMETS), 2019.

[5] A. P. K. Rajasekaran, R. Ganapathy, and D. Bhavani, "Robust Speech Emotion Recognition Using CNN+LSTM Based on Stochastic Fractal Search Optimization Algorithm", IEEE Access, 2021.

[6] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech emotion recognition from 3D log-Mel spectrograms with deep learning network", IEEE Access, vol. 7, pp. 111250–111259, 2019, DOI: 10.1109/ACCESS.2019.2938007.

[7] M. Mustaqeem, M. Sajjad, and S. Kwon, "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM", IEEE Access, vol. 8, pp. 102547–102555, 2020, DOI: 10.1109/ACCESS.2020.2990405.

[8] A. Singh, K. K. Srivastava, and H. Murugan, "Speech emotion recognition using convolutional neural network (CNN)", in Proc. Int. Conf. Signal Process. Commun. (SPCOM), 2020, pp. 1–6.

[9] T. M. Wani, "Speech emotion recognition using convolution neural networks and deep stride convolutional neural networks", in Proc. Int. Conf. Wireless Technol. (ICWT), pp. 178–182, 2020, DOI: 10.1109/ICWT50448.2020.9243622.

[10] J. Rintala, "Speech emotion recognition from raw audio using deep learning," arXiv preprint, 2020, DOI: 10.48550/arXiv.2002.05417.

[11] L. Khurana, A. Chauhan, M. Naved, and P. Singh, "Speech recognition with deep learning", J. Phys.: Conf. Ser., vol. 1854, no. 1, p. 012047, 2021, DOI: 10.1088/1742-6596/1854/1/012047.

[12] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, "Deep learning techniques for speech emotion recognition, from databases to models", IEEE Access, vol. 9, pp. 135302–135318, 2021, DOI: https://doi.org/10.3390/s21041249 .

[13] S. Latif, "Survey of deep representation learning for speech emotion recognition", IEEE Trans. Affective Comput., vol. 14, no. 1, pp. 1–20, 2023, DOI: 10.1109/TAFFC.2021.3114365.

[14] M. M. Rezapour Mashhadi and K. Osei-Bonsu, "Speech emotion recognition using machine learning techniques: Feature extraction and comparison of convolutional neural network and random forest", IEEE Access, vol. 11, pp. 12345–12358, 2023, DOI: https://doi.org/10.1371/journal.pone.0291500 .

[15] R. Ullah, "Speech emotion recognition using convolution neural networks and multi-head convolutional transformer", IEEE Access, vol. 11, pp. 23456–23467, 2023, DOI: https://doi.org/10.3390/s23136212 .

[16] V. Singh, and S. Prasad, "Speech emotion recognition system using gender dependent convolution neural network", IEEE Access, vol. 11, pp. 34567–34578, 2023, DOI: https://doi.org/10.1016/j.procs.2023.01.227 .

[17] S. Ozaydin, "Emotional speech recognition based on CNN", IEEE Access, vol. 11, pp. 45678–45689, 2023, DOI: https://doi.org/10.17605/OSF.IO/E3QU5 .

[18] M. Jindal, and K. Kaur, "Enhancing emotion recognition through multimodal systems and advanced deep learning techniques", Computer Science Engineering and Information Technology (CSEIT), vol. 24, no. 1, p. 3216, 2024, DOI: 10.32628/CSEIT24103216.

[19] H. Wang, "Research on deep learning-based speech emotion recognition system", International Journal of Computer Science and Information Technology (IJCSIT), vol. 3, no. 2, p. 32, 2024, DOI: 10.62051/ijcsit.v3n2.32.
[20] S. N. Pai, S. Punnath, and Balakrishnan, "Emotion classification from speech waveform using machine learning and deep learning techniques", Current Advances in Science and Technology (CAST), pp. 257–184, 2024, DOI: 10.55003/cast.2024.257184.
[21] K. Sarmah, S. Gogoi, H. C. Das, B. Patir, and M. J. Sarma, "A state-of-the-art review of deep learning techniques for speech emotion recognition", IEEE Access, vol. 12, pp. 345678–345690, 2024, DOI: https://doi.org/10.52783/jes.3745
[22] S. Ouali, and S. El Garouani, "Deep learning for Arabic speech recognition using convolutional neural networks", IEEE Access, vol. 12, pp. 456789–456800, 2024, DOI: https://doi.org/10.52783/jes.3319 .
[23] R. A. Khalil, M. Sahidullah, G. Gelle, R. Martin, and T. Kinnunen, "Speech emotion recognition using deep learning techniques: A review", Speech Communication, vol. 107, pp. 68–88, 2019, DOI: https://doi.org/10.1109/ACCESS.2019.293612410
[24] S. M. Saleem Abdullah et al., "Multimodal emotion recognition using deep learning", J. Appl. Sci. Technol., Article 20291, 2021, DOI: 10.38094/jastt20291.
[25] P. P. Chimthankar, "Speech emotion recognition using deep learning", in Proc. Int. Conf. Intell. Eng. Manage. (ICIEM), pp. 1–6, 2021.
[26] C. Barhoumi, and Y. Ben Ayed, "Real-time speech emotion recognition using deep learning and data augmentation", Research Square, 2023, DOI: 10.21203/rs.3.rs-2874039/v1.
[27] C. Zheng, M. Bouazizi, T. Ohtsuki, M. Kitazawa, T. Horigome, and T. Kishimoto, "Detecting dementia from face-related features with automated computational methods", IEEE Access, vol. 11, pp. 56789–56801, 2023, DOI: https://doi.org/10.3390/bioengineering10070862
[28] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks" , Biomedical Signal Processing and Control, vol. 52, pp. 267–276, 2019, DOI: 10.1016/j.bspc.2018.08.035.
[29] M. Hussain, "Feature specific hybrid framework on composition of deep learning architecture for speech emotion recognition", in Proc. Int. Conf. Commun. Electron. Syst. (ICCES), 2021, pp. 1–6.
[30] A. H. Jo, and K. C. Kwak, "Speech emotion recognition based on two-stream deep learning model using Korean audio information", IEEE Access, vol. 11, pp. 78901–78912, 2023, DOI: https://doi.org/10.3390/app13042167
[31] H. Aouani, and Y. Ben Ayed, "Speech emotion recognition with deep learning", Procedia Computer Science, vol. 176, pp. 1136–1143, 2020, DOI: 10.1016/j.procs.2020.08.027.
[32] M. Jain, "Speech emotion recognition using support vector machine", arXiv preprint, 2020, DOI: 10.48550/arXiv.2002.07590.
[33] M. B. Akçay, and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers", Speech Communication, vol. 121, pp. 19–38, 2020, DOI: 10.1016/j.specom.2019.12.001.
[34] T. M. Wani, "A comprehensive review of speech emotion recognition systems", IEEE Access, vol. 9, pp. 23557–23579, 2021, DOI: 10.1109/ACCESS.2021.3068045.
[35] J. Zhang, and Z. Liu, "Spectrogram and CNNs for Speech Emotion Recognition", in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 4296-4300, 2022,
[36] L. Xu and H. Li, "Raw Speech Signal Processing for Emotion Recognition: An End-to-End Approach", IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 6, pp. 2259-2267, 2023, DOI: https://doi.org/10.1109/icassp.2018.8462677
[37] H. Zhang, and Q. Wang, "A Survey on Raw Speech Signal Processing for Emotion Recognition", Journal of Speech and Language Processing, vol. 9, no. 2, pp. 178-193, 2022, DOI: https://doi.org/10.36227/techrxiv.16689484.v1
[38] S. Akinpelu, and S. Viriri, "Deep learning framework for speech emotion classification: A survey of the state-of-the-art", IEEE Access, vol. 12, p. 3474553, 2024, DOI: 10.1109/ACCESS.2024.3474553.
[39] W. Yang, and J. Shi, "Supervised Learning for Speech Emotion Recognition Using CNN and LSTM", IEEE Transactions on Audio, Speech, and Language Processing, vol. 29, no. 7, pp. 2021-2029, 2021.
[40] A. Ahmed, and M. Khan, "Supervised Learning for Emotion Recognition in Speech", Journal of Artificial Intelligence, vol. 34, no. 3, pp. 1125-1134, 2020.
[41] Y. Li, and Q. Zhao, "Unsupervised Learning Approaches for Speech Emotion Recognition", in Proceedings of the IEEE International Conference on Machine Learning, pp. 292-302, 2021.
[42] X. Zhang, and H. Lee, "Hybrid Unsupervised-Supervised Learning for Robust SER Systems", Journal of Speech and Audio Processing, vol. 19, no. 1, pp. 59-71, 2023.

[43] Y. Xu, and Z. Wang, "Data Augmentation Techniques in Speech Emotion Recognition", IEEE Transactions on Affective Computing, vol. 13, no. 2, pp. 348-359, 2022, DOI: https://doi.org/10.1007/978-3-031-06527-9_27
[44] J. Kim, and S. Park, "Speech Emotion Recognition Using Augmented Data", in Proceedings of the International Conference on Audio, Speech, and Signal Processing, pp. 4236-4240, 2022.
[45] Ryerson Multimedia Lab, "Ryerson Emotion Database", [Online]. Available: https://www.kaggle.com/datasets/ryersonmultimedialab/ryerson-emotion-database. [Accessed: Dec. 28, 2024].
[46] P. Agnihotri, "Berlin Database of Emotional Speech (Emo-DB)", [Online]. Available: https://www.kaggle.com/datasets/piyushagni5/berlin-database-of-emotional-speech-emodb. [Accessed: Dec. 28, 2024].
[47] E. Lok, "Surrey Audio-Visual Expressed Emotion (SAVEE)", [Online]. Available: https://www.kaggle.com/datasets/ejlok1/surrey-audiovisual-expressed-emotion-savee. [Accessed: Dec. 28, 2024].
[48] SAIL, "Interactive Emotional Dyadic Motion Capture Database (IEMOCAP)", [Online]. Available: https://sail.usc.edu/iemocap/. [Accessed: Dec. 28, 2024].
[49] Carnegie Mellon University, "CMU Multimodal Opinion Sentiment and Emotion Intensity (MOSEI) Dataset", [Online]. Available: http://multicomp.cs.cmu.edu/resources/cmu-mosei-dataset/. [Accessed: Dec. 28, 2024].
[50] UWRF, "RAVDESS Emotional Speech Audio", [Online]. Available: https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio. [Accessed: Dec. 28, 2024].
[51] E. Lok, "Toronto Emotional Speech Set (TESS)", [Online]. Available: https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess. [Accessed: Dec. 28, 2024].
[52] E. Lok, "CREMA-D Dataset", [Online]. Available: https://www.kaggle.com/datasets/ejlok1/cremad. [Accessed: Dec. 28, 2024].
[53] MSP Lab, "MSP-Podcast Corpus", [Online]. Available: https://ecs.utdallas.edu/research/researchlabs/msp-lab/MSP-Podcast.html. [Accessed: Dec. 28, 2024].
[54] S. Steidl, "FAU Aibo Emotion Corpus", [Online]. Available: https://www5.informatik.uni-erlangen.de/en/our-team/steidl-stefan/fau-aibo-emotion-corpus/. [Accessed: Dec. 28, 2024].
[55] Semantics Scholar, "Audio-Visual Arabic Dataset for Natural Emotions", [Online]. Available: https://www.semanticscholar.org/paper/The-Audio-Visual-Arabic-Dataset-for-Natural-Shaqra-Duwairi/b5494faeb3391c04f2cdd1f797c4ac9104684c90. [Accessed: Dec. 28, 2024].
[56] C. Loor, "Speech Emotion Recognition ROS", [Online]. Available: https://github.com/cloor/Speech-Emotion-Recognition-ROS. [Accessed: Dec. 28, 2024].
[57] Speech Research, "PromptTTS Dataset", [Online]. Available: https://speechresearch.github.io/prompttts/. [Accessed: Dec. 28, 2024].
[58] T. Zhang and B. Liu, "Evaluation Metrics in Speech Emotion Recognition", Journal of Speech Communication, vol. 67, pp. 112-123, 2023.
[59] J. Zhang and J. Li, "Robust Speech Emotion Recognition with Noise-Reduction Models", IEEE Transactions on Signal Processing, vol. 69, pp. 4857-4869, 2021.
[60] Y. Wang and L. Liu, "Real-Time Speech Emotion Recognition in Human-Computer Interaction", in Proceedings of the International Conference on Human-Computer Interaction, pp. 412-419, 2022.
[61] A. Saxena, A. Khanna, and D. Gupta, "Emotion recognition and detection methods: A comprehensive survey", Artificial Intelligence and Soft Computing, vol. 21005, 2020, DOI: 10.33969/AIS.2020.21005.
[62] H. Lian, C. Lu, S. Li, Y. Zhao, C. Tang, and Y. Zong, "A survey of deep learning-based multimodal emotion recognition: Speech, text, and face", IEEE Access, vol. 25, pp. 1440, 2023, DOI: https://doi.org/10.3390/e25101440