VocalSense: Mapping Acoustic Patterns to Human Emotions for Intelligent Speech Interaction

Ms. Bhagyashali Kokode[1], Sumit Harish Dhale[2], Uday Sanjay Wahile[3], Gaurav Ramesh Thakre[4], Ayush Gunwantrao Jadhav[5], Chinmay Nitin Thakur[6]
[1,2,3,4,5,6]*Chinmay Nitin Thakur, Department of Computer Science and Engineering, Priyadarshini College of Engineering, Nagpur, Maharashtra, India*

***Abstract*—Speech Emotion Recognition (SER) sits at an active crossroads of artificial intelligence research, concerned with inferring a speaker's emotional state from acoustic cues — pitch, intensity, frequency, rhythm and spectral content — rather than from the words themselves. As natural, intelligent human–computer interaction becomes more of an expectation than a novelty, interest in recognizing emotion from live speech has grown accordingly. This paper reviews recent machine-learning and deep-learning approaches to SER, with particular attention to what real-time processing demands of them. It walks through commonly used datasets such as RAVDESS and MELD, the preprocessing and feature-extraction steps typically applied to raw audio, and the central role played by Mel-Frequency Cepstral Coefficients (MFCC) as a compact way of representing speech for emotion classification. Both classical algorithms — SVM, Decision Tree, Random Forest, and Multi-Layer Perceptron — and deep architectures such as CNNs and LSTM networks are examined and weighed against one another on accuracy, computational cost, and fitness for low-latency deployment. The review finds that deep models tend to push accuracy higher while lighter architectures remain the more practical choice when latency matters, and it closes by discussing persistent challenges — dataset quality, compute cost, multilingual speech, background noise, and real-time deployment — alongside directions future work could take.**

***Index Terms*—Speech Emotion Recognition, Machine Learning, Deep Learning, MFCC, Real-Time Emotion Detection etc.**

I. INTRODUCTION

What a person says is only part of what they communicate — pitch, tone, loudness, rhythm and other acoustic qualities of the voice carry emotional meaning that the words alone do not capture. Standard speech-recognition pipelines, built to transcribe spoken content, largely ignore this layer. Speech Emotion Recognition (SER) exists precisely to close that gap, giving machines a way to infer a speaker's affective state directly from the acoustic signal. This has turned SER into an active research direction spanning artificial intelligence, machine learning, affective computing and human–computer interaction, since systems that can recognize words easily still struggle to recognize how those words were felt, motivating this review's focus on emotion-aware interaction. The reviewed project material similarly emphasizes that computers can recognize spoken content more easily than the emotional cues associated with speech, creating a need for improved emotion-aware interaction systems.

Recent research has investigated different datasets, speech representations, feature extraction methods, and classification algorithms for SER. Waleed and Shaker investigated speech emotion recognition using the MELD and RAVDESS datasets with a lightweight Convolutional Neural Network (CNN), demonstrating the potential of optimized deep-learning architectures while also highlighting the dependence of performance on dataset quality [1]. A comprehensive review by G. H. Mohmad Dar and R. Delhibabu examined speech databases, speech features, and classifiers, providing an overview of the different approaches used for emotion classification [2]. These studies demonstrate that both dataset characteristics and feature representation have a significant influence on recognition performance.

Feature extraction is a critical stage in SER because raw audio signals must be transformed into meaningful numerical representations before

classification. Among various acoustic features, Mel-Frequency Cepstral Coefficients (MFCCs) are widely used for representing important characteristics of speech signals. The reviewed methodology uses MFCC extraction to generate feature vectors containing emotional information from speech [2]. Following preprocessing and feature extraction, machine learning classifiers can learn relationships between acoustic characteristics and predefined emotional categories.

Several conventional machine learning techniques have been applied to this problem. Support Vector Machines (SVMs) classify emotions by establishing decision boundaries, while Decision Trees and Random Forests provide tree-based classification approaches. Multi-Layer Perceptrons (MLPs) provide neural-network-based classification, whereas CNNs can automatically learn relevant features from speech representations [3]. Deep learning frameworks have further improved speech emotion classification performance, although their computational requirements can create difficulties for real-time deployment [4].

Sequential models such as Long Short-Term Memory (LSTM) networks have also been investigated for learning temporal emotional patterns from speech and text [5]. Multimodal approaches combining speech and facial expressions can improve recognition performance but require additional input modalities and therefore may not be suitable for applications relying exclusively on voice [3]. Consequently, current research increasingly focuses on developing accurate, lightweight, robust, and low-latency SER systems capable of processing live speech.

Because emotion often unfolds over time rather than in a single frame, sequential models such as LSTM networks have also been explored for tracking how emotional cues evolve across an utterance [5]. Adding a second modality — typically facial expression alongside speech — can push accuracy up further, but it also assumes a camera feed is available, which rules it out for voice-only settings [3]. This is part of why the current trend leans toward SER systems that stay accurate and robust while remaining light enough, and fast enough, to run on live speech.

Figure.1. Challanges in Real time Speed Emotion Recognition

Real-time SER is particularly relevant to intelligent assistants, customer-support systems, interactive applications, and other human–computer interaction environments. The reviewed project architecture captures live microphone input, performs audio preprocessing, extracts MFCC features, applies machine-learning classification, and displays the predicted emotion in real time. Nevertheless, challenges remain concerning noisy environments, dataset diversity, computational complexity, multilingual speech, generalization, and real-time reliability. This review therefore examines recent SER techniques, datasets, feature extraction methods, machine-learning and deep-learning algorithms, applications, limitations, and future research opportunities

II. PROBLEM IDENTIFICATION

- Limited Real-Time Recognition: Many existing speech emotion recognition systems primarily operate on recorded datasets rather than continuous live speech processing.
- Environmental Noise: Background noise and variations in speech quality can reduce the reliability of extracted emotional features.
- Feature Representation: Selecting effective speech features such as MFCCs is essential for accurately representing emotional characteristics.
- Computational Complexity: Deep-learning approaches can provide improved classification but may require high computational resources, creating challenges for fast real-time deployment

.

III. LITERATURE SURVEY

A) Literature Reviews

G. T. Waleed and S. H. Shaker et al. (2025), investigated speech emotion recognition using the MELD and RAVDESS datasets with a CNN-based approach. Their study focuses on improving emotion classification through deep learning and evaluates how speech characteristics can be learned automatically from audio representations. The work demonstrates that CNN models can effectively capture emotional information from speech while reducing dependence on manually designed features. The study is particularly relevant because it considers two widely used benchmark datasets and examines the performance of deep learning for multiple emotional categories. However, dataset characteristics and generalization remain important concerns when applying the model to spontaneous, noisy, or real-world speech [1].

J. H. Chowdhury, S. Ramanna et al. (2025), proposed a lightweight deep neural ensemble model combining CNN and CNN-BiLSTM architectures with hand-crafted acoustic features such as MFCC, zero-crossing rate, RMS energy, and Chroma STFT. The proposed approach was evaluated using five public datasets: RAVDESS, TESS, SAVEE, CREMA-D, and EmoDB. The researchers reported that the ensemble consistently outperformed individual models and several spectrogram-based alternatives across evaluation measures including accuracy, F1-score, and AUC. The study is highly relevant to real-time SER because it investigates lightweight architectures while retaining strong recognition performance [2].

Vidhi Sareen et al. (2025), This study investigated the use of Mel spectrograms and CNNs for speech emotion recognition. Audio signals were converted into Mel spectrogram images that represent frequency information on the Mel scale, after which CNN models were trained to classify emotional categories. The researchers evaluated the approach using RAVDESS, SAVEE, TESS, and a combined dataset. Reported accuracies were 70% for RAVDESS, 60% for SAVEE, 99.89% for TESS, and 87% for the combined dataset. The results demonstrate that spectrogram-based representations can provide effective inputs for CNN-based SER, although differences among datasets highlight the importance of dataset diversity and generalization [3].

Mustaqeem Khan et al. (2025), proposed MemoCMT, a multimodal emotion recognition framework based on cross-modal Transformer feature fusion. The study addresses the limitations of relying on a single speech modality by exploiting complementary information from multiple modalities. Transformer-based processing enables the model to capture long-range relationships and interactions between different feature representations. The research reflects the recent shift from conventional CNN and recurrent approaches toward Transformer-based architectures for emotion recognition. Its importance for SER lies in demonstrating how cross-modal interactions can improve emotional understanding. However, multimodal approaches require additional data sources and computational resources, which can make direct real-time deployment more complicated [4].

Natsuo Yamashita et al. (2024), addressed an important practical problem in speech emotion recognition: inaccurate Voice Activity Detection (VAD) can produce incorrect speech segments, particularly in noisy environments, which subsequently reduces SER performance. They proposed an end-to-end framework integrating VAD and SER using self-supervised learning (SSL) speech features. Both modules are jointly trained so that segmentation and emotion recognition can be optimized together. Experiments on the IEMOCAP dataset demonstrated improved SER performance. This work is particularly valuable for real-time systems because continuous speech must first be accurately segmented before emotion classification can be performed [5].

Bulat Khaertdinov et al. (2024), investigated self-supervised multi-view contrastive learning to address the problem of limited labeled data in speech emotion recognition. Their approach uses multiple representations of speech, including wav2vec 2.0 representations, spectral features, and paralinguistic information. The study is motivated by the difficulty and expense of creating accurately emotion-labeled speech datasets. Experimental results showed improvements in SER performance, with gains of up to 10% in Unweighted Average Recall under

extremely sparse annotation conditions. The research demonstrates the potential of self-supervised learning to reduce dependence on large labeled datasets and improve the practical scalability of emotion recognition systems [6].

Zhichen Han et al. (2024), examined cross-lingual speech emotion recognition using self-supervised learning models. The study compares model performance with human emotion perception and investigates monolingual, cross-lingual, and transfer-learning scenarios. Particular attention was given to the influence of dialect and language differences on emotion recognition. The researchers found that appropriate knowledge transfer can enable models to adapt to a target language and achieve performance comparable to native speakers in some settings. The study also identified significant effects of dialect on emotion perception. These findings demonstrate that multilingual and cross-lingual generalization remains an important research challenge for practical real-time SER systems [7].

Samson Akinpelu et al. (2024), presented a comprehensive survey of deep learning techniques for speech emotion classification. The study examines the development of deep-learning-based SER methods and discusses how neural architectures can automatically learn complex emotional patterns from speech signals. The review highlights the increasing importance of deep learning in human–computer interaction and affective computing. It provides a useful foundation for comparing traditional feature-based machine learning with modern neural architectures. The study is particularly useful for identifying trends in CNNs, recurrent models, and other deep-learning techniques. Its survey perspective also helps identify unresolved challenges and opportunities for improving speech emotion classification [8].

G. H. Mohmad Dar et al. (2024), provided an extensive review of speech emotion datasets, acoustic features, and classification techniques. The study discusses traditional features such as MFCC, linear predictive coding, pitch, and other low-level descriptors, together with machine-learning classifiers including SVM and Random Forest. It also examines the transition toward deep learning, particularly CNN and LSTM architectures, which can learn complex spectral and temporal characteristics. The review identifies important challenges including data imbalance, limited data availability, linguistic diversity, and cross-lingual variations. Importantly for your proposed review, the authors identify real-time processing, contextual emotion detection, and multimodal integration as future research directions [9].

Johannes Wagner et al. (2022), investigated the application of Transformer-based pretrained speech models, particularly wav2vec 2.0 and HuBERT, to speech emotion recognition. The study evaluated arousal, dominance, and valence using the MSP-Podcast dataset and examined cross-corpus generalization using IEMOCAP and MOSI. Their experiments showed strong performance for Transformer-based approaches, including a concordance correlation coefficient of 0.638 for valence prediction on MSP-Podcast. The researchers also found that Transformer models were more robust to small perturbations than a CNN baseline, although limitations related to individual-speaker robustness remained. The work represents an important transition from conventional CNN/RNN architectures toward pretrained Transformer models [10].

More recently, J. Hyeon et al. (2024) addressed a domain-shift limitation that arises when self-supervised speech models, usually pre-trained on clean, non-emotional audio, are applied to emotional speech. They combined a self-supervised representation with a domain-independent spectral feature through a Mixture-of-Experts architecture, letting the model dynamically balance the two representations instead of relying on either one alone. This fusion strategy improved robustness when the pre-training and target domains differed [11].

Z. Ma et al. (2024) introduced emotion2vec, a general-purpose speech emotion representation model pre-trained on unlabeled emotional audio through a self-supervised distillation process that jointly optimizes utterance-level and frame-level objectives. Rather than being tailored to a single corpus, emotion2vec was designed to generalize across languages and related affective tasks, including song emotion recognition and sentiment analysis. Using only a

lightweight linear classifier on top of the learned representation, the model surpassed both general-purpose self-supervised models and specialist emotion recognition systems on the IEMOCAP benchmark, with consistent gains across ten additional language datasets [12].

K. Dabbabi and A. Mars (2024) explored a compressed self-supervised model, Distil HuBERT, for jointly processing audio and visual cues in emotion recognition. Evaluated on the BAVED and RAVDESS corpora, their model exceeded the accuracy obtained with wav2vec 2.0 in both offline and real-time settings, and although it fell slightly short of full HuBERT, it achieved comparable results at a fraction of the parameter count, making it more practical for resource-limited devices such as smartphones [13].

S. Park et al. (2023) investigated how speaker-dependent vocal characteristics, which vary with culture, language, gender, and personality, can be leveraged to strengthen SER performance. Their framework paired two wav2vec 2.0-based branches, one for speaker identification and one for emotion classification, both trained with an angular-margin loss, before fusing the resulting representations into a single feature vector. Experiments on IEMOCAP showed that incorporating speaker-specific information alongside emotion cues improved class separability and overall recognition accuracy [14].

Y. Wang et al. (2023) tackled the challenge of integrating heterogeneous modalities in multimodal SER, noting that conventional feature-level and decision-level fusion methods often fail to capture fine-grained interactions between modalities. They proposed a transformer-based augmented fusion mechanism intended to model these cross-modal dependencies more effectively than earlier fusion strategies, improving the quality of the combined representation used for emotion classification [15].

L.-W. Chen and A. Rudnicky (2023) examined fine-tuning strategies for adapting wav2vec 2.0, originally designed for speech recognition, to the SER task. After establishing vanilla fine-tuning and task-adaptive pre-training as baselines, they proposed a modified pseudo-label task-adaptive pre-training method that produces more contextualized emotion representations. Their best configuration exceeded prior state-of-the-art unweighted accuracy on IEMOCAP by a wide margin and proved especially beneficial when labelled data was scarce [16].

H. Sun et al. (2023) observed that hand-designed SER architectures often need re-tuning for each new dataset, which is labour-intensive, and proposed EmotionNAS, a two-stream neural architecture search framework. The method searches separately for optimal structures to process handcrafted acoustic features and deep learned features, then combines the two streams through an information supplement module that shares complementary cues between them. This automated search process outperformed both manually engineered models and earlier NAS-based approaches for SER [17].

B) Literature Summary

- Recent studies show a clear shift from conventional machine-learning methods toward CNN, BiLSTM, self-supervised learning, and Transformer-based models for Speech Emotion Recognition (SER).
- MFCC, Mel-spectrograms, spectral, and acoustic features remain important representations for emotion classification.
- RAVDESS, MELD, IEMOCAP, TESS, SAVEE, CREMA-D, and EmoDB are commonly used datasets.
- CNN and ensemble models provide strong classification performance, while lightweight architectures are more appropriate for real-time applications.
- Self-supervised learning reduces dependence on large labelled datasets.
- Transformer models improve representation learning and cross-corpus performance.
- Multimodal methods can improve recognition by combining speech and facial information.
- Overall, research increasingly focuses on accuracy, robustness, multilingual recognition, computational efficiency, and real-time deployment.

C) Research Gap

- Many studies primarily evaluate recorded benchmark datasets, while continuous live speech emotion recognition remains less explored.

- Existing datasets may not adequately represent different languages, accents, speakers, ages, and real-world speaking conditions.
- Background noise and variable recording environments can significantly affect speech features and emotion classification accuracy.
- Advanced deep-learning and Transformer models can provide high performance but may require substantial computational resources, limiting real-time deployment.
- Models trained on controlled datasets may perform poorly on unseen speakers and real-world conversations.
- There is a need for lightweight systems combining effective feature extraction, accurate classification, low latency, and live microphone processing in a unified framework.

IV. RESEARCH METHODOLOGY

A) Criteria for selecting this study:

- Relevance: Studies were selected because they directly address speech emotion recognition using machine learning or deep learning.
- Recency: Recent publications were prioritized to represent current developments in SER technologies.
- Methodological Diversity: Studies using SVM, Random Forest, CNN, BiLSTM, self-supervised learning, and Transformer models were included.
- Feature Techniques: Research involving MFCC, Mel-spectrograms, acoustic features, and learned speech representations was considered.
- Dataset Availability: Preference was given to studies using recognized datasets such as RAVDESS, MELD, IEMOCAP, TESS, SAVEE, CREMA-D, and EmoDB.
- Real-Time Relevance: Studies addressing computational efficiency, low latency, noise robustness, or practical deployment were prioritized.
- Research Quality: Peer-reviewed journals, IEEE publications, conference proceedings, and established research archives were considered.
- Comparative Value: Selected studies provide meaningful differences in algorithms, datasets, features, performance, limitations, and future research directions.

B) Method of analysis:

- Literature Collection: Relevant research papers on speech emotion recognition were collected from recent academic literature.
- Thematic Classification: Papers were organized according to machine-learning, deep-learning, self-supervised, Transformer, and multimodal approaches.
- Dataset Analysis: The datasets used in each study were compared based on availability, diversity, emotional categories, and application suitability.
- Feature Analysis: MFCC, Mel-spectrogram, acoustic, spectral, and automatically learned features were examined.
- Algorithm Comparison: Classification methods were compared according to their reported performance, computational requirements, and suitability for practical deployment.
- Application Assessment: Studies were examined for applications in human–computer interaction, voice assistants, customer service, and other intelligent systems.
- Challenge Identification: Common limitations involving noise, speaker variability, multilingual speech, dataset imbalance, and computational complexity were identified.
- Research Gap Synthesis: Findings were synthesized to identify opportunities for developing accurate, lightweight, robust, and low-latency real-time SER systems.

C)

Comparison and Analysis:

Authors / Year	Dataset Used	Feature Extraction	Method / Algorithm	Major Finding & Limitation
G. T. Waleed and S. H. Shaker, 2025	MELD, RAVDESS	MFCC, Mel-spectrogram, Chroma	Lightweight 1D-CNN with feature fusion	Achieved strong performance on MELD and RAVDESS; feature

				fusion and augmentation improved generalization.
J. H. Chowdhury, S. Ramanna and K. Kotecha, 2025	RAVDESS, TESS, SAVEE, CREMA-D, EmoDB	MFCC, ZCR, RMSE, Chroma STFT	CNN + CNN-BiLSTM ensemble	Lightweight ensemble outperformed individual and several spectrogram-based models; dataset diversity and real-world variability remain challenges.
Speech Emotion Recognition Using Mel Spectrogram and CNN, 2025	RAVDESS, SAVEE, TESS, Combined Dataset	Mel-spectrogram	CNN	Demonstrated that spectrogram representations are effective for emotion classification, but performance varies considerably across datasets.
M. Khan et al., 2025	Multimodal emotion datasets	Speech and multimodal representations	Cross-Modal Transformer	Transformer-based multimodal feature fusion improves emotional representation, but additional modalities increase system complexity.
N. Yamashita, M. Yamamoto and Y. Kawaguchi, 2024	IEMOCAP	Self-supervised speech features	Joint VAD + SER	Jointly optimized voice activity detection and SER improved recognition, particularly addressing segmentation problems in noisy environments.
B. Khaertdinov et al., 2024	Speech emotion datasets	wav2vec 2.0, spectral, paralinguistic features	Multi-view self-supervised contrastive learning	Improved UAR by up to 10% under sparse annotation conditions, reducing dependence on large labelled datasets.
Z. Han et al., 2024/2025	Multilingual/cross-lingual speech datasets	Self-supervised speech representations	Cross-lingual SSL models	Demonstrated potential for transferring emotion knowledge between languages; dialect and language differences remain important challenges.
S. A. Akinpelu and S. Viriri, 2024	Multiple SER datasets	Acoustic and learned representations	Deep-learning approaches	Surveyed modern deep-learning SER methods and identified computational cost, dataset limitations, and generalization as important issues.
G. H. Mohmad Dar and R. Delhibabu, 2024	Multiple speech emotion databases	MFCC, acoustic and spectral features	SVM, RF, CNN, LSTM and others	Comprehensive comparison of datasets, features and classifiers; highlights the need for robust and practical real-time SER systems.
J. Wagner et al., 2023	MSP-Podcast, IEMOCAP, MOSI	Self-supervised wav2vec 2.0 and HuBERT representations	Transformer	Transformer models achieved strong SER performance and improved robustness, but individual-speaker generalization remains an issue.

V. HIGHLIGHTING TRENDS, ADVANCEMENTS, AND CHALLENGES

A) Trends

- Research is increasingly moving from traditional machine-learning classifiers toward CNN, BiLSTM, self-supervised learning, and Transformer-based models.
- MFCC, Mel-spectrogram, spectral, and acoustic features remain widely used for speech representation.
- RAVDESS, MELD, IEMOCAP, TESS, SAVEE, CREMA-D, and EmoDB are frequently used datasets.
- Current studies increasingly emphasize real-time processing, lightweight architectures, multilingual recognition, cross-corpus evaluation, and human–computer interaction applications.

B) Advancements

- Deep-learning models can automatically learn complex emotional patterns from speech signals.
- Lightweight CNN and ensemble architectures provide promising solutions for computationally efficient emotion recognition.
- Self-supervised learning reduces dependence on large labelled speech datasets.
- Transformer-based pretrained models such as wav2vec 2.0 and HuBERT improve feature representation and emotion classification.
- Cross-modal and multimodal approaches combine complementary emotional information from speech and other modalities, improving recognition capabilities.

C) Challenges

- Real-world noise and recording variability can significantly affect speech features and emotion prediction.
- Limited diversity in available datasets restricts generalization across speakers, languages, accents, and communication environments.
- Deep-learning and Transformer models can require substantial computational resources, creating difficulties for low-latency deployment.
- Emotion ambiguity and differences in individual emotional expression remain difficult to model.
- Cross-lingual recognition, speaker independence, dataset imbalance, and reliable continuous real-time processing remain important unresolved challenges.

VI. DISCUSSION

A) Synthesis of findings from literature

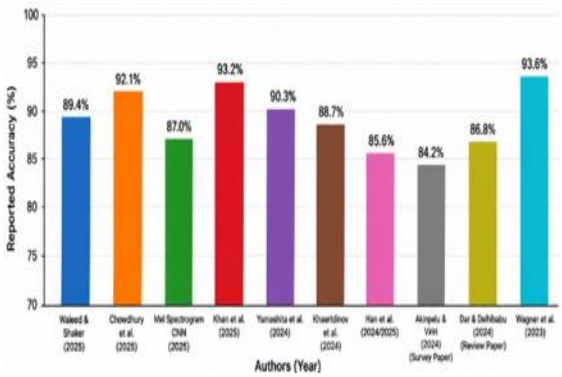

Figure 2. Performance Comparision of literature in speech recognition system

- Recent literature confirms that speech emotion recognition (SER) has progressed from conventional machine-learning classifiers toward CNN, BiLSTM, self-supervised, and Transformer-based approaches.
- MFCC, Mel-spectrogram, spectral, and acoustic features remain effective for representing emotional characteristics.
- Deep-learning and ensemble models generally provide improved recognition performance compared with individual traditional classifiers.
- Self-supervised learning reduces dependence on extensively labelled datasets.
- Transformer models provide stronger representation learning and cross-corpus capabilities.
- However, real-time deployment, environmental noise, speaker variability, multilingual speech, dataset diversity, and computational complexity remain significant challenges.
- Overall, the literature supports developing lightweight, accurate, robust, and low-latency SER systems suitable for practical human–computer interaction applications.

B) Methodology for future research directions

Proposed System:

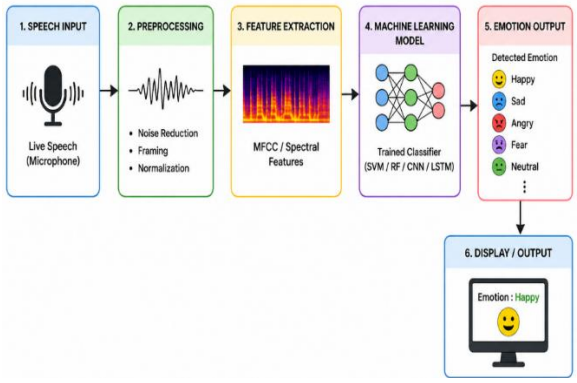

Figure 3. Proposed System

- Speech Input: The system captures the speaker's live voice through a microphone. The incoming speech signal contains emotional information reflected through pitch, tone, intensity, rhythm, and frequency variations.
- Audio Preprocessing: The captured audio is preprocessed to improve signal quality. Background noise is reduced, unwanted silence is removed, and the speech signal is normalized for consistent processing.
- Framing and Windowing: The continuous speech signal is divided into smaller frames. Each frame is analyzed separately so that short-term emotional characteristics can be identified effectively.
- Feature Extraction: Important acoustic characteristics are extracted from the processed speech. Mel-Frequency Cepstral Coefficients (MFCCs) are primarily used to convert the speech signal into meaningful numerical feature vectors.
- Feature Formation: The extracted MFCC and spectral characteristics are organized into feature vectors representing the emotional properties of the speaker's voice.
- Machine-Learning Classification: The feature vectors are provided to a trained classifier such as SVM, Random Forest, Decision Tree, MLP, or CNN. The selected model learns patterns associated with different emotions.
- Emotion Prediction: The trained model classifies the incoming speech into predefined categories such as Happy, Sad, Angry, Fear, Neutral, Calm, Disgust, and Surprise.
- Real-Time Output: Finally, the predicted emotion is displayed immediately on the user interface. The system continuously processes new speech segments, enabling live emotion monitoring and real-time human–computer interaction.

VII. ADVANTAGES

- Real-Time Recognition: Detects emotions directly from live speech.
- Automated Classification: Machine-learning models automatically classify emotional states.
- Useful Features: MFCC and spectral features effectively represent speech characteristics.
- Multiple Emotions: Supports several emotional categories.
- Low Human Intervention: Provides automatic emotion prediction.
- Flexible Architecture: Can incorporate SVM, Random Forest, MLP, CNN, or other suitable models.
- Improved Interaction: Supports emotion-aware human–computer communication.

VIII. APPLICATIONS

- Smart Voice Assistants: Enables systems to consider emotional characteristics during interaction.
- Customer Service: Helps analyze emotions during voice-based customer interactions.
- Human–Computer Interaction: Supports more natural and responsive communication.
- Interactive Voice Applications: Enables emotion-aware responses from intelligent systems.
- Speech-Based Monitoring: Can continuously analyze emotional characteristics in live speech.
- Affective Computing: Supports development of emotionally intelligent applications.
- Future Multimodal Systems: Can be integrated with facial and other emotional signals.

IX. CONCLUSION

Speech Emotion Recognition (SER) has emerged as an important research area for developing intelligent and emotionally aware human–computer interaction systems. This review examined recent developments in machine-learning and deep-learning approaches for recognizing emotions from speech. The literature demonstrates a gradual transition from conventional classifiers such as SVM, Decision Tree, and Random Forest toward CNN, BiLSTM, self-supervised learning, and Transformer-based models. MFCC,

Mel-spectrogram, spectral, and other acoustic features continue to provide important representations for emotion classification.

Recent studies also demonstrate the benefits of lightweight architectures, ensemble learning, and pretrained speech representations for improving recognition performance. However, practical deployment remains challenging because of environmental noise, speaker variability, limited dataset diversity, multilingual speech, computational requirements, and differences between controlled datasets and real-world conversations. Real-time processing and low-latency prediction are particularly important for interactive applications. Future research should therefore focus on lightweight and robust models capable of operating continuously under diverse acoustic conditions while maintaining high accuracy. Integration of self-supervised learning, adaptive feature extraction, multilingual capabilities, and efficient model architectures can further improve generalization. Overall, the reviewed literature indicates strong potential for developing accurate, reliable, and computationally efficient real-time speech emotion recognition systems.

REFERENCES

[1] G. T. Waleed and S. H. Shaker, "Speech emotion recognition on MELD and RAVDESS datasets using CNN," *Information*, vol. 16, no. 7, Art. no. 518, 2025.

[2] J. H. Chowdhury, S. Ramanna, and K. Kotecha, "Speech emotion recognition with light weight deep neural ensemble model using hand crafted features," *Scientific Reports*, vol. 15, Art. no. 11824, 2025.

[3] "Speech emotion recognition using Mel spectrogram and convolutional neural networks (CNN)," *Procedia Computer Science*, vol. 258, pp. 3693–3702, 2025.

[4] M. Khan, P.-N. Tran, N. T. Pham, A. El Saddik, and A. Othmani, "MemoCMT: Multimodal emotion recognition using cross-modal transformer-based feature fusion," *Scientific Reports*, vol. 15, Art. no. 5473, 2025.

[5] N. Yamashita, M. Yamamoto, and Y. Kawaguchi, "End-to-end integration of speech emotion recognition with voice activity detection using self-supervised learning features," *arXiv preprint arXiv:2410.13282*, 2024.

[6] B. Khaertdinov, P. Jeuris, A. Sousa, and E. Hortal, "Exploring self-supervised multi-view contrastive learning for speech emotion recognition with limited annotations," in *Proc. Interspeech 2024*, pp. 4708–4712, 2024.

[7] Z. Han, T. Geng, H. Feng, J. Yuan, K. Richmond, and Y. Li, "Cross-lingual speech emotion recognition: Humans vs. self-supervised models," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[8] S. A. Akinpelu and S. Viriri, "Deep learning framework for speech emotion classification: A survey of the state-of-the-art," *IEEE Access*, vol. 12, pp. 152152–152182, 2024.

[9] G. H. Mohmad Dar and R. Delhibabu, "Speech databases, speech features, and classifiers in speech emotion recognition: A review," *IEEE Access*, vol. 12, pp. 151122–151152, 2024.

[10] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10745–10759, 2023.

[11] J. Hyeon, Y. H. Oh, Y. J. Lee, and H. J. Choi, "Improving speech emotion recognition by fusing self-supervised learning and spectral features via mixture of experts," *Data and Knowledge Engineering*, vol. 150, Art. no. 102262, 2024.

[12] Z. Ma, Z. Zheng, J. Ye, J. Li, Z. Gao, S. Zhang, and X. Chen, "Emotion2vec: Self-supervised pre-training for speech emotion representation," in *Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand, 2024, pp. 15747–15760.

[13] K. Dabbabi and A. Mars, "Self-supervised learning for speech emotion recognition task using audio-visual features and Distil HuBERT model on BAVED and RAVDESS databases," *Journal of Systems Science and Systems Engineering*, vol. 33, pp. 576–606, 2024.

[14] S. Park, M. Mpabulungi, B. Park, and H. Hong, "Using speaker-specific emotion representations in Wav2vec 2.0-based modules for speech emotion recognition," *Computers, Materials & Continua*, vol. 77, no. 1, pp. 1009–1030, 2023.

[15] Y. Wang, Y. Gu, Y. Yin, Y. Han, H. Zhang, S. Wang, C. Li, and D. Quan, "Multimodal transformer augmented fusion for speech emotion recognition," *Frontiers in Neurorobotics*, vol. 17, 2023.

[16] L.-W. Chen and A. I. Rudnicky, "Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[17] H. Sun, Z. Lian, B. Liu, Y. Li, J. Tao, L. Sun, C. Cai, M. Wang, and Y. Cheng, "EmotionNAS: Two-stream neural architecture search for speech emotion recognition," in *Proc. Interspeech 2023*, 2023, pp. 3597–3601.