Speech and Audio Processing: A Comprehensive Survey on Speech Emotion Recognition (SER)

KARTHIKEYAN.K
ECE-A

Abstract

Speech Emotion Recognition (SER) plays a critical role in human-computer interaction, allowing systems to understand and respond to users' emotional states. This paper presents a survey of recent advancements in SER, covering studies published between 2023 and 2025. We analyze various methodologies, including deep learning models, multimodal approaches, and feature extraction techniques. Key challenges such as cross-lingual robustness, real-time processing, and security vulnerabilities are identified. Our comparative analysis highlights current trends and provides future research directions to enhance SER effectiveness and applicability.

.

I. INTRODUCTION

The growing integration of artificial intelligence in everyday applications has led to an increased focus on understanding human emotions. SER aims to bridge the gap between humans and machines by enabling systems to recognize and interpret emotions from speech. Applications range from customer service to mental health monitoring and assistive technologies. Despite recent advancements, challenges such as dataset biases, multilingual support, and computational complexity persist. This survey aims to provide a comprehensive review of SER methodologies, trends, and future research directions.

II. BACKGROUND AND SIGNIFICANCE

SER is increasingly used in various domains, including:

- **Healthcare**: For diagnosing mental health conditions, emotional well-being monitoring, and rehabilitation.
- **Customer Service**: To enhance user interactions, automate responses, and improve user experience.
- **Education**: For student engagement analysis, adaptive learning, and personalized education.
- **Human-Robot Interaction**: Enhancing the effectiveness of communication between machines and humans.
- **Security and Forensics**: Emotion recognition for lie detection, criminal investigations, and security surveillance.
- **Entertainment**: Emotion-driven adaptive gaming and content recommendations.

Emotion representation in speech involves prosody, pitch, tone, speech patterns, and contextual dependencies, making SER a complex and multi-faceted problem.

III. METHODOLOGY

To conduct this survey, we systematically searched IEEE Xplore, Springer, Elsevier, and other reputed databases for research articles published between 2023 and 2025. Selection criteria included relevance to SER, methodological contributions, and impact. We categorized papers based on their focus, such as feature extraction, deep learning models, multimodal approaches, and security considerations.

IV. DATASETS USED IN SER RESEARCH

Major datasets used for SER include:

- **IEMOCAP** (Interactive Emotional Dyadic Motion Capture Database)
- **RAVDESS** (Ryerson Audio-Visual Database of Emotional Speech and Song)
- **EmoDB** (Berlin Database of Emotional Speech)
- **CREMA-D** (Crowd-Sourced Emotional Multimodal Actors Dataset)
- **MELD** (Multimodal Emotion Lines Dataset)
- **MSP-Podcast**: A dataset extracted from real-world podcast recordings.
- **SAVEE**: A database focusing on emotion-labelled speech collected from male actors.

V. REVIEW OF EXISTING WORK

Recent advancements in SER can be classified into the following categories:

A. Feature Extraction Techniques

Feature extraction is a crucial step in SER, influencing system accuracy. Common features include:

- **Mel-Frequency Cepstral Coefficients (MFCCs)**: Captures spectral information.
- **Spectrograms**: Time-frequency representation of speech.
- **Prosodic Features**: Includes pitch, energy, and duration.
- **Wavelet Transforms**: Extracts features at multiple resolutions.

- **Deep Feature Representations**: Leveraging deep neural networks for automatic feature extraction.
- **Voice Activity Detection (VAD)**: Enhancing signal-to-noise ratio for improved recognition.
- **Phoneme-based Features**: Considering phonetic structures to enhance emotion discrimination.

B. Deep Learning Models

Neural networks, including CNNs, RNNs, and Transformers, have significantly improved SER performance. Advanced techniques include:

- **Attention Mechanisms**: Enhancing focus on crucial speech segments.
- **Self-Supervised Learning**: Leveraging unlabeled data for training.
- **Graph Neural Networks (GNNs)**: Capturing interdependencies between speech segments.
- **Contrastive Learning**: Improving emotion differentiation.
- **Recurrent Transformers**: Combining RNNs and Transformers for sequential data processing.
- **Diffusion Models**: Leveraging generative modelling for emotional feature enhancement.

C. Multimodal Approaches

Integrating speech with facial expressions and physiological signals enhances emotion recognition accuracy. Recent studies have explored multimodal fusion strategies, including:

- **Audio-Visual Fusion**: Combining speech with facial expressions.
- **Physiological Signal Integration**: Including EEG and heart rate data.
- **Cross-Modal Attention Mechanisms**: Improving robustness across modalities.
- **Wearable Sensor Data**: Integration with real-time physiological sensing devices.

VI. CHALLENGES IN REAL-WORLD SER APPLICATIONS

Real-world SER applications face challenges such as:

- **Background Noise**: Reducing interference from environmental noise.
- **Speaker Variability**: Adapting to different accents and speech patterns.
- **Emotion Ambiguity**: Addressing overlapping emotional states.
- **Model Interpretability**: Enhancing transparency of SER decisions.
- **Low-Resource Languages**: Expanding SER models for multilingual support.

- **Real-time Deployment**: Optimizing SER models for embedded and edge devices.
- **Scalability**: Ensuring SER models can handle large-scale user interactions.

TABLE I
COMPARATIVE ANALYSIS OF RECENT SER STUDIES

Study	*Focus Area*	*Methodology*	*Key Findings*	*Limitations*
Dai et al. (2025)	*Feature Extraction*	*Combined softmax crossentropy loss with center loss*	*Improved accuracy*	*Limited to spectrogram inputs*
Shen et al. (2024)	*Classification Models*	*Emotion Neural Transducer (ENT)*	*Outperformed state-of-theart methods*	*Requires low word error rate*
Wu et al. (2023)	*Integrated Systems*	*Jointly-trained system (AER, ASR, SD)*	*Enhanced performance*	*System complexity*
Lu et al. (2023)	*Feature Aggregation*	*Local to Global Feature Aggregation (LGFA)*	*Superior performance*	*Computationally intensive*
Gurowiec & Nissim (2024)	*Security Aspects*	*Cyber-attack analysis on SER systems*	*Identified 70% attacks unaddressed*	*No mitigation strategies*

- ## VII. DISCUSSION & FUTURE SCOPE

- Future research should focus on:
- **Developing robust cross-lingual SER systems.**
- **Enhancing real-time processing efficiency.**
- **Addressing security vulnerabilities in SER systems.**
- **Expanding datasets to cover diverse emotional expressions.**
- **Investigating federated learning for privacy-preserving SER.**
- **Hybrid SER models combining symbolic AI with deep learning.**
- **Emotion Transition Modeling**: Understanding the evolution of emotions over time.
- **Zero-Shot Learning for SER**: Enabling recognition of unseen emotions.

VIII. CONCLUSION

This survey highlights advancements in SER research, categorizing methodologies and identifying research gaps. Future efforts should prioritize cross-lingual adaptability, security, and real-time processing to improve SER applications. Addressing these challenges will contribute to more effective and ethical human-computer interactions.