International Journal of Advance and Innovative Research
Volume 12, Issue 2: April - June 2025
ISSN 2394 - 7780

ADVANCING SPEECH EMOTION RECOGNITION: DEEP LEARNING ARCHITECTURES, DATASET CONSIDERATIONS, AND ROBUST EVALUATION METHODOLOGIES

[1]Irfan Chaugule and [2]Dr. Satish R Sankaye
[1]Research Scholar, MGM University, DR.G.Y. Pathrikar College of Computer Science and Information Technology, Chhatrapati Sambhajinagar, Maharashtra
irfanchaugule@gmail.com

[2]MGM University, DR.G.Y. Pathrikar College of Computer Science and Information Technology, Chhatrapati Sambhajinagar, Maharashtra
Sankayesr@gmail.com

ABSTRACT
Speech Emotion Recognition (SER) aims to enable machines to understand human emotional states from vocal expressions, a capability crucial for intelligent human-computer interaction. The rapid advancement of deep learning (DL) has significantly propelled the SER field. This paper provides a comprehensive review of key DL architectures employed in SER, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (LSTMs, GRUs), hybrid models, and Transformer-based approaches. It critically examines commonly used emotional speech datasets, highlighting their characteristics, inherent limitations, and the crucial distinctions between acted, elicited, and spontaneous emotions, alongside challenges in cross-corpus generalization. Furthermore, standard evaluation metrics and protocols essential for robust SER research are detailed. Building upon this review, the paper outlines a framework for the systematic development and evaluation of SER systems, emphasizing rigorous experimental design. The goal is to foster best practices in model selection, data handling, and evaluation, thereby contributing to the development of more accurate, reliable, and generalizable SER technologies.

Keywords: *Speech Emotion Recognition, Deep Learning, CNN, LSTM, GRU, Transformers, Emotional Speech Datasets, Evaluation Metrics, Cross-Corpus Evaluation, Research Methodology.*

1. INTRODUCTION: NAVIGATING THE LANDSCAPE OF DEEP LEARNING FOR SPEECH EMOTION RECOGNITION
The human voice is a rich conduit of emotional information, extending far beyond the literal meaning of words.[2] Harnessing this information through Speech Emotion Recognition (SER) has become a focal point of research, promising to revolutionize human-computer interaction, mental health diagnostics, customer service, and more.[2] Deep learning (DL) has been instrumental in this endeavor, with models like Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, Gated Recurrent Units (GRUs), and Transformers demonstrating remarkable capabilities in deciphering complex emotional patterns from speech.[2]

However, the path to effective SER systems is multifaceted. Beyond the crucial stage of data preprocessing, the selection of appropriate DL architectures, the careful consideration of available emotional speech datasets, and the application of rigorous evaluation methodologies are paramount. Each DL architecture possesses unique strengths in capturing different aspects of speech signals—CNNs excel at local spectro-temporal patterns, while RNNs model temporal dependencies, and Transformers offer powerful contextual understanding through attention mechanisms.[2] The datasets used for training and evaluation significantly shape model performance and generalizability, with critical differences between acted, elicited, and spontaneous emotional expressions posing substantial challenges.[19, 20, 21, 2] Furthermore, robust evaluation requires appropriate metrics that account for issues like class imbalance, and protocols that ensure speaker independence and assess cross-corpus generalization.[2]

This paper aims to provide a comprehensive overview of these critical components. It delves into the prominent DL architectures used in SER, profiles key emotional speech datasets discussing their limitations, and outlines standard evaluation practices. Building on this, it proposes a framework for the systematic development and evaluation of SER systems, with the objective of fostering evidence-based practices and advancing the creation of more effective and reliable SER technologies.

2. FUNDAMENTALS OF SPEECH EMOTION RECOGNITION
A Speech Emotion Recognition (SER) system typically follows a pipeline structure to transform raw audio input into a predicted emotional state. [2] The core stages include:

1. **Speech Input:** The initial raw audio signal.

ISSN 2394 - 7780

2. **Preprocessing:** Cleaning the audio and preparing it for feature extraction (e.g., noise reduction, normalization).
3. **Feature Extraction:** Deriving informative parameters (features) from the speech signal.
4. **Data Augmentation:** Artificially expanding the training dataset.
5. **Classification/Regression Model:** The deep learning model that learns to map features to emotions.
6. **Emotion Output:** The predicted emotional state.

This paper focuses on the **Classification/Regression Model** stage, the **Datasets** used to train and evaluate these models, and the **Evaluation Metrics and Protocols** essential for assessing their performance.

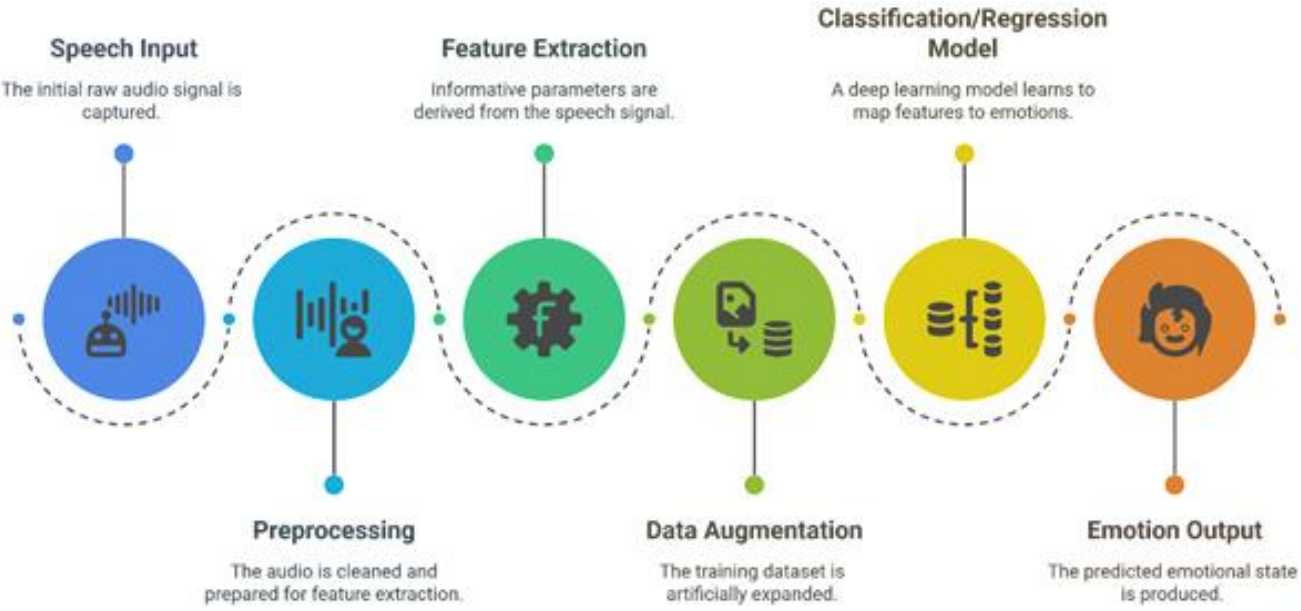

Figure 1: Fundamentals of Speech Emotion Recognition

3. DEEP LEARNING ARCHITECTURES IN SER: AN IN-DEPTH REVIEW

Deep learning has become the predominant approach in SER due to its ability to automatically learn hierarchical and discriminative feature representations from complex speech data.[2]

3.1. Convolutional Neural Networks (CNNs)

CNNs are highly effective for processing grid-like data, such as 2D Mel-spectrograms derived from speech. [2]

- **Architecture:** Typically comprise convolutional layers (to detect local patterns like spectral shapes), non-linear activation functions (e.g., ReLU), pooling layers (to reduce dimensionality and provide invariance), and fully connected layers for classification. [2] Early layers learn simple local features, while deeper layers learn more complex, abstract patterns. [2]
- **Input:** Most commonly, 2D Mel-spectrograms. 1D CNNs applied to raw audio or 1D features (e.g., MFCCs) are also explored. [2] Some studies suggest 1D CNNs combined with LSTMs can outperform standalone 2D CNNs by effectively capturing temporal patterns. [22, 23, 2]
- **Strengths:** Excel at extracting salient local spectro-temporal features, robust to input variations, and model spatial/spectral aspects well. [2]

3.2. Recurrent Neural Networks (RNNs) – LSTM and GRU

RNNs are designed for sequential data, making them ideal for modeling the temporal dynamics of speech and emotional expressions. [2]

3.2.1. Long Short-Term Memory (LSTM) Networks

LSTMs use memory cells and gating mechanisms (forget, input, output gates) to regulate information flow, enabling them to capture long-range dependencies and mitigate vanishing/exploding gradient problems. [24, 25, 2] They process sequences of frame-wise features (e.g., MFCCs) to model dynamic changes indicative of emotion. [2]

ISSN 2394 - 7780

3.2.2. Gated Recurrent Unit (GRU) Networks

GRUs simplify the LSTM architecture with fewer parameters (update and reset gates), often achieving comparable performance with potentially faster training. [26, 27, 2] Overfitting can be an issue, addressed by techniques like dropout and batch normalization. [26, 27, 2]

3.2.3. Bidirectional RNNs (Bi-LSTM, Bi-GRU)

These process input sequences in both forward and backward directions, providing access to past and future context simultaneously, which is beneficial as emotional cues can appear late in an utterance. [28, 2]

- **Input for RNNs:** Typically sequences of frame-wise acoustic features like MFCCs. [2]
- **Strengths of RNNs:** Excellent at modeling temporal dependencies and contextual information crucial for dynamic emotional expressions. [2]

3.3. Hybrid Models (e.g., CNN-LSTM, CNN-GRU)

Hybrid models combine CNNs (for robust local feature extraction from spectrograms) and RNNs (to model temporal sequences of these features). [2]

- **Architecture:** A common setup involves a CNN front-end processing spectrograms, with its output feature maps fed into an LSTM or GRU backend to capture temporal context before classification. [2]
- **Input:** Usually Mel-spectrograms for the CNN part. [2]
- **Strengths:** Capture both fine-grained spectro-temporal details (CNNs) and long-term temporal dependencies (RNNs), often achieving state-of-the-art performance. [2] Zhao et al. reported a CNN-LSTM achieving 52.14% unweighted accuracy on IEMOCAP. [2]

3.4. Attention Mechanisms and Transformers

3.4.1. Attention Mechanisms

Attention mechanisms allow DL models to dynamically weigh the importance of different parts of the input sequence or feature map, focusing on the most emotionally salient segments. [2] They have been successfully integrated into CNNs, LSTMs, and hybrid models.

3.4.2. Transformer Models

Transformers rely entirely on self-attention mechanisms, allowing for more parallelization and effective modeling of very long-range dependencies compared to RNNs. [29, 30, 2] The self-attention mechanism weighs the importance of all other frames in a sequence when encoding a particular frame. [30, 2] Multi-head attention runs this process in parallel with different projections, capturing diverse relational aspects. [30, 2]

3.4.3. Pre-trained Transformer Models for Speech (Foundation Models)

Large-scale Transformer models pre-trained on vast amounts of unlabeled speech (e.g., Wav2Vec2.0, HuBERT) using self-supervised learning can be fine-tuned for SER, often achieving state-of-the-art results. [31, 32, 2]

- **Wav2Vec2.0:** Learns contextualized speech representations from raw audio by predicting masked parts of the input. [31, 32, 2] Fine-tuned versions are powerful feature extractors for SER. [31, 32, 2]
- **HuBERT (Hidden-Unit BERT):** Uses an offline clustering step (e.g., k-means on MFCCs) to generate pseudo-labels (hidden units) and trains to predict these for masked regions, addressing challenges like overlapping sound units and variable lengths without explicit segmentation. [33, 34, 2]

Table 1: Key Characteristics of Deep Learning Architectures for SER

Architecture	Core Principle	Typical Inputs	Strengths for SER	Weaknesses/ Challenges	Example References (from Snippets)
CNN	Hierarchical learning of local patterns using convolutional filters. [2]	Mel-spectrograms (2D), MFCCs/raw audio (1D). [2]	Extracts salient local spectro-temporal features; Robust to input	Limited ability to model long-range temporal dependencies if not very	[22, 23, 2]

ISSN 2394 - 7780

			variations; Good for image-like data. [2]	deep or combined with other mechanisms.	
LSTM	Gated recurrent units (input, forget, output gates) to model long-range temporal dependencies. [24, 2]	Sequences of features (e.g., MFCCs). [2]	Excellent at modeling temporal dynamics and context; Mitigates vanishing/exploding gradients. [2]	Can be computationally intensive; May overfit on small datasets if not regularized.	[24, 25, 2]
GRU	Simplified gated recurrent units (update, reset gates) for temporal modeling. [26, 27, 2]	Sequences of features (e.g., MFCCs). [2]	Similar to LSTM in performance, often faster training and fewer parameters. [2]	Can still overfit; Performance relative to LSTM can be task-dependent.	[26, 27, 2]
CNN-LSTM (Hybrid)	CNN for spatial/local feature extraction, LSTM for temporal modeling of CNN outputs. [2]	Mel-spectrograms. [2]	Combines strengths: captures fine-grained spectro-temporal details and long-term dependencies. [2]	More complex architecture; Requires careful design of CNN-RNN interface.	[2]
Transformer	Self-attention mechanisms to weigh importance of all parts of a sequence simultaneously. [30, 2]	Raw audio (Wav2Vec2), sequences of features. [31, 2]	Excellent at very long-range dependencies; Highly parallelizable; Foundation models (Wav2Vec2, HuBERT) learn powerful general representations from unlabeled data. [31, 30, 33, 34, 2]	Requires large datasets for pre-training foundation models; Can be computationally demanding; Still evolving for SER.	[29, 31, 30, 32, 33, 34, 2]

ISSN 2394 - 7780

4. EMOTIONAL SPEECH DATASETS: LANDSCAPE, CHALLENGES, AND SELECTION

High-quality, well-annotated emotional speech datasets are fundamental for training and evaluating SER models. [2] However, their creation and use present challenges like limited size, difficulty in capturing genuine emotions, and scarcity for languages other than English. [19, 20, 2] Real-world data also exhibit variability not always present in lab recordings. [35, 2]

4.1. In-depth Profiles of Common Datasets

The choice of dataset significantly influences SER model development and generalizability. [2]

Table 2: Overview of Commonly Used Speech Emotion Recognition Datasets

Dataset Name	Language	#Speakers (M/F)	#Utterances (Approx.)	Emotions	Recording Type	Key Features	Limitations	Original Reference(s)
IEMOCAP	English	10 (5M/5F)	~10,000 turns (~12 hrs)	Happy, sad, angry, neutral, frustrated, excited, surprise, fear, disgust (9 total); also valence, arousal, dominance	Mixed (Scripted & Spontaneous dyadic interactions)	Multimodal (audio, video, motion capture), naturalistic dyadic interactions, categorical & dimensional labels.	Moderate inter-evaluator agreement for some labels; elicited emotions though with spontaneous segments; actors performing scripts. [19, 36, 37, 20, 2]	Busso et al. (2008) [36, 37, 2]
RAVDESS	English (N. American)	24 (12M/12F)	7,356 files (1,440 speech files)	Speech: Calm, happy, sad, angry, fearful, surprise, disgust (7 emotions + neutral). Song: Calm, happy, sad, angry, fearful.	Acted	High-quality audio-visual, 2 intensity levels (normal, strong), lexically-matched statements, balanced gender. [38, 39, 2]	Acted emotions (may lack subtlety); controlled environment. [20, 2]	Livingstone & Russo (2018) [39, 2]
EMODB (Berlin)	German	10 (5M/5F)	535 utterances	Anger, boredom, disgust, fear, happy, sad, neutral (7 emotions)	Acted	High emotional quality, clear expressions, anechoic chamber recording. [40, 41, 2]	Relatively small size, single language (German), acted emotions. [40, 2]	Burkhardt et al. (2005) [40, 41, 2]
SAVEE	English (British)	4 (4M/0F)	480 utterances	Anger, disgust, fear, happy, sad, surprise, neutral (7 emotions)	Acted	Audio-visual, TIMIT sentences used. [42, 2]	Very small number of speakers, all male (potential bias), acted emotions. [42, 2, 43]	Haq & Jackson (2009) [42, 2]

ISSN 2394 - 7780

TESS	English (N. American)	2 (0M/2F)	2,800 samples	Anger, disgust, fear, happy, sad, surprise, neutral (7 emotions)	Acted	Clear emotional expressions by two actresses, carrier phrase "Say the word ___". [44, 45, 46, 2, 47, 48]	Very limited speaker variability (2 female speakers only), acted emotions, potentially overly simplistic. [49, 46, 2]	Pichora-Fuller & Dupuis (2010) [44, 45, 2]
CREMA-D	English	91 (48M/43F)	7,442 clips	Happy, sad, angry, fearful, disgust, neutral (6 emotions)	Acted	Diverse actors (ethnicity, age), multiple modalities, varying emotional intensity levels, crowd-sourced annotations. [50, 51, 2]	Acted emotions, surprise not included, modest human recognition for audio-only (40.9%). [50, 51, 2]	Cao et al. (2014) [50, 51, 2]

- **IEMOCAP:** ~12 hours' audiovisual data, 10 actors, dyadic interactions (scripted & improvised), categorical & dimensional labels. Limitation: moderate inter-evaluator agreement, elicited context. [19, 36, 37, 20, 2]
- **RAVDESS:** 24 actors, 2 lexically-matched statements, 7 speech emotions (+ neutral), 2 intensities. High-quality audio-visual. Limitation: acted emotions. [38, 39, 2]
- **EMODB:** German, 10 actors, 535 utterances, 7 emotions. High acoustic quality (anechoic). Limitations: small, single language, acted. [40, 41, 2]
- **SAVEE:** 4 male British English speakers, 480 utterances (TIMIT sentences), 7 emotions. Limitations: very few speakers, all male, acted. [42, 2, 43]
- **TESS:** 2 actresses, 2800 audio files ("Say the word ___"), 7 emotions. Limitations: extremely limited speaker variability, acted. [44, 45, 49, 46, 2, 47, 48]
- **CREMA-D:** 91 diverse actors, 7442 clips, 6 emotions, varying intensities, crowd-sourced labels. Limitation: acted, modest human audio-only recognition (40.9%). [50, 51, 2]

4.2. Acted vs. Spontaneous vs. Elicited Emotions: The Realism Dilemma

Emotional datasets are categorized by how emotions were obtained [10, 20, 2]:

- **Acted (Simulated):** Actors portray emotions (e.g., RAVDESS, EMODB). Advantages: controlled, clear, high-quality audio, easier labeling, often higher model accuracy. Disadvantages: can be stereotypical, exaggerated, may not reflect genuine emotion; models may perform poorly on spontaneous data. [19, 42, 21, 2] Accuracy on acted SAVEE (78.75%) vs. spontaneous IEMOCAP (50.06%) highlights this gap. [21, 2]
- **Elicited (Induced):** Emotions induced via stimuli (e.g., IEMOCAP partially). Advantages: more naturalistic than acted. Disadvantages: induction effectiveness varies, ethical considerations. [19, 20, 2]
- **Spontaneous (Natural):** Genuine emotions in real-life situations (e.g., call centers). Advantages: most ecologically valid. Disadvantages: hard to collect/annotate, noisy, lower inter-annotator agreement, scarce, models perform worse. [20, 2]

This leads to a "realism-performance trade-off." Robust performance on spontaneous data is more meaningful for practical utility. [2]

ISSN 2394 - 7780

4.3. Cross-Corpus Evaluation and Domain Adaptation

Models trained on one corpus often perform poorly on another (cross-corpus generalization problem) due to mismatches in languages, accents, recording conditions, speaker demographics, emotion styles, and annotation schemes. [52, 53, 54, 28, 2]

Approaches to address this (Domain Adaptation) include [2]:

- **Feature Normalization:** Corpus-level, speaker-level normalization. [52, 2]
- **Transfer Learning:** Instance weighting, subspace learning (e.g., TNNMF, AKTLR [52, 2]), adversarial training, multi-task learning (e.g., emotion and gender recognition with subdomain adaptation [53, 54, 2]), adapter modules for pre-trained models [55, 2], and Parameter-Efficient Fine-Tuning (PEFT) for large models like Wav2Vec2/HuBERT (e.g., two-stage adaptation from acted to natural emotions). [56, 2]

Success in cross-corpus evaluation is crucial for practical SER systems. [2]

5. Evaluation Metrics and Protocols in SER

Rigorous evaluation requires appropriate metrics and standardized protocols.

5.1. Standard Performance Metrics

Metrics must account for imbalanced emotion classes. [20, 2]

- **Accuracy:** Correct predictions / total predictions. Misleading for imbalanced data. [2]
- **Precision:** TP/(TP+FP). Accuracy of positive predictions for a class. [2]
- **Recall (Sensitivity):** TP/(TP+FN). Ability to identify all positive instances of a class. [2]
- **F1-score:** 2×(Precision×Recall)/(Precision+Recall). Harmonic mean, balances precision and recall. Macro F1 (unweighted average per class) and Weighted F1 (weighted by class support) are used. [2]
- **Unweighted Average Recall (UAR) / Mean Recall:** Average of recall scores for each class. Gives equal importance to each class, robust for imbalanced tasks. [20, 2]
- **Confusion Matrix:** Visualizes performance, showing correct classifications and misclassifications between classes. [2]

For imbalanced SER, UAR and Macro F1-score are generally more informative than raw accuracy. [2]

5.2. Evaluation Protocols

- **Speaker-Independent Evaluation:** Ensures speakers in training set do not appear in validation/test sets. Crucial for generalization to unseen speakers. Methods: Leave-One-Speaker-Out (LOSO) cross-validation, fixed speaker splits. [2]
- **Cross-Corpus Evaluation:** Training on one dataset and testing on another. Rigorously tests generalization across different conditions. [2]
- **Data Splitting:** Clearly defined train/validation/test splits, consistently used. Validation set for hyperparameter tuning.
- **Statistical Significance Testing:** Use tests like paired t-tests or ANOVA ($p<0.05$) to determine if performance differences are statistically significant. [2]

Adherence to rigorous protocols is fundamental for advancing SER. [2]

6. A Framework for Systematic Development and Evaluation of SER Systems

Building on the reviewed architectures, datasets, and metrics, a systematic framework is essential for developing and evaluating SER systems. This adapts the blueprint from [2] with a focus on model and dataset considerations.

6.1. Overall Research Design Philosophy

The framework emphasizes [2]:

1. **Baseline Establishment:** Define baseline performance for each model-dataset pair with minimal preprocessing.
2. **Systematic Model Comparison:** Evaluate different DL architectures using consistent preprocessing and evaluation metrics.

ISSN 2394 - 7780

3. **Dataset Impact Analysis:** Assess model performance across diverse datasets to understand data-dependent effects and generalization.
4. **Rigorous Evaluation:** Employ speaker-independent and cross-corpus protocols with appropriate metrics.
5. **Statistical Validation:** Confirm the significance of performance differences.

6.2. Rationale for Selection of Deep Learning Architectures
The evaluation should span representative DL architectures to understand their suitability for SER [2]:

- **M1 (CNN):** 2D CNN for Mel-spectrograms (local spectro-temporal patterns).
- **M2 (LSTM):** Bi-LSTM for sequential features (long-range temporal dependencies).
- **M3 (GRU):** Bi-GRU as an alternative to LSTM.
- **M4 (CNN-LSTM Hybrid):** Combines CNN feature extraction with LSTM temporal modeling.
- Consideration of Transformers: While the initial blueprint focused on the above, evaluations should increasingly include Transformer-based models (e.g., fine-tuning Wav2Vec2, HuBERT) due to their state-of-the-art potential.
  Consistent hyperparameter tuning or settings are needed for fair comparisons.

6.3. Rationale for Dataset Selection and Splitting Strategy
To ensure robust findings, evaluations must use multiple diverse datasets [2]:

- **Selection Criteria:** Vary recording conditions (acted, spontaneous), language, speaker numbers, emotion categories (e.g., RAVDESS, IEMOCAP, EMODB).
- **Splitting Strategy:** Enforce standardized speaker-independent train/validation/test splits. Harmonize emotion categories for comparison.

6.4. Experimental Setup and Implementation Choices
Consistency is key for valid comparisons [2]:

- **Software/Libraries:** Standard tools (Python, Librosa, PyTorch/TensorFlow, Scikit-learn).
- **Training Parameters:** Consistent optimizer, loss function, batch size, epochs with early stopping, and weight initialization when comparing models or dataset impacts.
- **Hardware:** GPUs for feasible training.
- **Reproducibility:** Fix random seeds, document parameters, share code/configurations.

6.5. Rigorous Evaluation Protocol Design
Reinforce the use of appropriate metrics and protocols [2]:

- **Primary Metrics:** UAR and Macro F1-score.
- **Secondary Analyses:** Per-class metrics, confusion matrices.
- **Evaluation Types:** Strict speaker-independent and cross-corpus evaluations.
- **Statistical Significance:** Validate all comparative claims.

7. Discussion: Model Efficacy, Dataset Utility, and Methodological Rigor in SER
The systematic application of the framework allows for a deeper understanding of SER system development.

Interpreting Model Efficacy:

- **CNNs** are effective for spectrogram-based local feature learning. The choice of 1D vs. 2D CNNs and their combination with RNNs can influence performance based on how temporal information is captured. [22, 23, 2]
- **LSTMs and GRUs** excel at modeling temporal sequences from frame-based features like MFCCs. Bidirectional variants often provide an advantage by using broader context. [2]
- **Hybrid CNN-RNN models** often achieve strong performance by combining spatial feature extraction with temporal modeling. [2]
- **Transformers and Pre-Trained Models (Wav2Vec2, HuBERT)** represent the cutting edge, learning powerful representations from vast amounts of data. Their ability to model long-range dependencies and be

ISSN 2394 - 7780

fine-tuned for SER is a significant advantage, often reducing the need for extensive manual feature engineering or complex augmentation. [31, 32, 56, 2] Parameter-Efficient Fine-Tuning (PEFT) makes adapting these large models more feasible. [56, 2]

Dataset Utility and Challenges:

- The "acted vs. spontaneous" dilemma remains a core challenge. While acted datasets (RAVDESS, EMODB) yield higher accuracies due to clearer expressions, models trained on them often fail to generalize to real-world spontaneous emotions. [20, 21, 2] Datasets like IEMOCAP offer a mix but still have limitations. [36, 37, 2]
- Dataset size and diversity (speakers, languages, recording conditions) are critical. Many existing datasets are small or limited in these aspects. [10, 19, 35, 20, 2]
- Cross-corpus generalization is poor due to feature distribution mismatches. Domain adaptation techniques are crucial but challenging to perfect. [52, 53, 54, 28, 2] Techniques incorporating SER-specific acoustic knowledge or sophisticated adaptation of foundation models show promise. [52, 56, 2]

Methodological Rigor:

- The lack of standardized evaluation protocols across studies hinders direct comparison and slows progress. Adherence to speaker-independent evaluation is a minimum requirement.
- Metrics like UAR and Macro F1-score are essential for imbalanced data but not universally adopted over simpler accuracy.
- Statistical validation of results is often overlooked but necessary for robust claims.

Implications for SER System Design:

- **Model Selection:** Choice depends on data availability, computational resources, and desired performance. Pre-trained Transformers are powerful but may require significant resources for fine-tuning. Hybrid models offer a good balance for many scenarios.
- **Data Strategy:** Prioritize diverse, naturalistic data if possible. If using acted data, be aware of generalization limits and explore domain adaptation.
- **Robust Evaluation:** Always use speaker-independent splits and appropriate metrics. Cross-corpus evaluation is a vital stress test.

8. CONCLUSION AND FUTURE DIRECTIONS IN SER SYSTEMS AND EVALUATION

This paper has reviewed the landscape of deep learning architectures, emotional speech datasets, and evaluation methodologies critical for advancing Speech Emotion Recognition. The journey from raw speech to recognized emotion is complex, with model choice, data characteristics, and evaluation rigor playing pivotal roles.

DL models, from CNNs and LSTMs to sophisticated Transformers, offer powerful tools for SER. However, their effectiveness is deeply intertwined with the datasets they are trained on. The field grapples with the limitations of existing datasets, particularly the gap between acted and spontaneous emotions, and the challenge of cross-corpus generalization. Robust and standardized evaluation practices are essential for navigating these challenges and fostering genuine progress.

Future directions in SER systems and evaluation methodologies include:

- **Advancements in DL Architectures:** Continued exploration of self-supervised learning, novel Transformer variants, and architectures that better integrate contextual and paralinguistic information for SER.
- **Creation of More Realistic Datasets:** Development of larger, more diverse datasets featuring spontaneous, "in-the-wild" emotional speech across various languages and cultures, with careful attention to annotation quality and ethics.
- **Improved Domain Adaptation Techniques:** More effective and SER-specific domain adaptation methods to bridge the gap between different corpora and conditions, particularly for leveraging pre-trained foundation models.
- **Standardization of Evaluation Protocols:** Wider adoption of common, rigorous evaluation protocols, including standardized dataset splits and reporting of comprehensive metrics, to improve comparability across studies.

ISSN 2394 - 7780

- **Fairness, Accountability, and Transparency (FAT) in SER:** Investigating and mitigating biases in models and datasets related to demographic attributes (gender, age, accent, culture), and developing more interpretable SER systems.
- **Multimodal SER:** Better integration of speech with other modalities (text, video, physiological signals) for more robust and nuanced emotion understanding.
- **Low-Resource SER:** Developing techniques that perform well with limited labeled data, crucial for many languages and specific emotional contexts.

By addressing these future directions with a commitment to methodological rigor, the SER community can build more accurate, reliable, and ethically sound systems capable of truly understanding the emotional depth of human communication.

9. REFERENCES

[1] Vary, P. (1985). Noise suppression by spectral magnitude estimation—Mechanism and theoretical limits. Signal Processing, 8(4), 387-400. [1] / Berouti, M., Schwartz, R., & Makhoul, J. (1979). Enhancement of speech corrupted by acoustic noise. ICASSP '79. [1]

[2] (Internal Document: Optimizing Data Preprocessing Pipelines for Enhanced Speech Emotion Recognition Using Deep Learning)

[3] Boll, S. F. (1979). Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustics, Speech, and Signal Processing, 27(2), 113-120. (as per content of [2] referring to snippet [1], and [3] itself)

[4] Wiener, N. (1949). Extrapolation, Interpolation, and Smoothing of Stationary Time Series. MIT Press. [4]

[5] Vaseghi, S.V. (2000). Advanced Digital Signal Processing and Noise Reduction. John Wiley & Sons.

[6] Islam, M. R., et al. (2025). A Comprehensive Review of Deep Learning Approaches for Speech Enhancement. Algorithms, 18(5), 272.

[7] Yechuri, S., & Vanabathina, S. D. (2025). Speech Enhancement: A Review of Different Deep Learning Methods. International Journal of Neural Systems.

[8] Codecademy. (2025). Normalization. Codecademy Articles.

[9] DeepLearning.AI Community. (2020). Confusing on normalisation.

[10] Nwe, T. L., et al. (2003). Speech emotion recognition without explicit segmentation. Oriental COCOSDA. (General SER features)

[11] Feng, E., et al. (2024). Probing Handcrafted Acoustic Features from Pre-trained Speech Embeddings for Emotion Recognition. arXiv preprint arXiv:2409.09511.

[12] Eyben, F., et al. (2016). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. IEEE Transactions on Affective Computing, 7(2), 190-202. [12]

[13] Eskimez, S. E., et al. (2022). Perceptual Fine-Tuning of Speech Enhancement Models with Differentiable eGeMAPS Features. arXiv preprint arXiv:2207.00237.

[14] Eyben, F., et al. (2016). [12]

[15] Park, D. S., et al. (2019). SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. arXiv preprint arXiv:1904.08779.

[16] Park, D. S., et al. (2019). SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. Proceedings of Interspeech 2019.

[17] Ronneberger, F., et al. (2024). Generating Mel-Spectrograms with GANs for Data Augmentation in Industrial Sound Analysis. Proceedings of DAS-DAGA 2025.

[18] Lin, Y.-H., et al. (2024). Voiceprint Recognition Enhancement via Generative Adversarial Networks and Optimal Data Balancing. Algorithms, 17(12), 583.

[19] Geminiani, J. C. B., et al. (2024). EMOVOME: A New Spontaneous Multimodal Dataset for Speech Emotion Recognition in Spanish with User Validation. arXiv preprint arXiv:2403.02167v1.

[20] Geminiani, J. C. B., et al. (2024). EMOVOME: A New Spontaneous Multimodal Dataset for Speech Emotion Recognition in Spanish with User Validation. arXiv preprint arXiv:2403.02167v3.

[21] Chenchah, F., & Lachiri, Z. (2014). Speech Emotion Recognition in Acted and Spontaneous Context. Procedia Computer Science, 39, 139-145.

[22] Kumar, S. A., et al. (2023). Human–Computer Interaction with a Real-Time Speech Emotion Recognition with Ensembling Techniques, 1D Convolution Neural Network, and Attention. Applied Sciences, 13(3), 1829.

[23] Kumar, S. A., et al. (2023). Human–Computer Interaction with a Real-Time Speech Emotion Recognition with Ensembling Techniques, 1D Convolution Neural Network, and Attention. Applied Sciences, 13(3), 1829. [22]

[24] Barman, I. R., & Shanthini, A. (2023). Speech Emotion Recognition Using Deep Learning Algorithm. In Advanced Practical Approaches to Deep Learning Models. River Publishers.

[25] ResearchGate. (n.d.). LSTM-based architecture for speech emotion recognition.

[26] Delantar, R. A. A., et al. (2024). An Enhancement of Gated Recurrent Unit (GRU) for Speech Emotion Recognition in the Implementation of Voice-Based Danger Recognition System. UIJRT, 6(3). [27]

[27] Delantar, R. A. A., et al. (2024). An Enhancement of Gated Recurrent Unit (GRU) for Speech Emotion Recognition in the Implementation of Voice-Based Danger Recognition System. UIJRT, 6(3).

[28] Sefik, I. E., & Cernocky, J. (2019). A Cross-Corpus Study on Speech Emotion Recognition. Text, Speech, and Dialogue (TSD 2019).

[29] Reddit. (2023). [Article] Recognition of Emotion in Speech-related Audio Files with LSTM-Transformer. r/Scholar.

[30] Latif, S., et al. (2023). Transformers in Speech Processing: A Survey. arXiv preprint arXiv:2303.11607. [30]

[31] Wang, N., & Yang, D. (2025). Speech emotion recognition using fine-tuned Wav2vec2.0 and neural controlled differential equations classifier. PLoS ONE, 20(2), e0318297.

[32] Pepino, L., Riera, P., & Ferrer, L. (2021). Emotion Recognition from Speech Using wav2vec 2.0 Embeddings. Proceedings of Interspeech 2021.

[33] UnrealSpeech Blog. (2024). Exploring HuBERT: A Revolutionary Approach to Self-Supervised Speech Representation Learning.

[34] GeeksforGeeks. (n.d.). HuBERT Model.

[35] Singh, A., et al. (2025). A Survey on Emotion Recognition and Generation using Deep Learning. arXiv preprint arXiv:2502.06803v1.

[36] Busso, C., et al. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42(4), 335-359.

[37] Busso, C., et al. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42(4), 335-359. [36]

[38] Ordóñez-Rivas, D. E., et al. (2024). Developing a Dataset of Audio Features to Classify Emotions in Speech. Applied Sciences, 13(2), 39. [39]

[39] Liu, Y., et al. (2025). A Deep Learning Model for Speech Emotion Recognition Integrating CNN and Transformer with Cross-Attention Fusion. The Journal of Supercomputing, 16(5), XXX-XXX. (RAVDESS creation details).

[40] Papers With Code. (n.d.). EmoDB Dataset.

[41] Wagner, J., et al. (2023). Dawn of the Transformer Era in Speech Emotion Recognition: A Survey. arXiv preprint arXiv:2312.06270.

[42] Haq, S., & Jackson, P. J. B. (2009). Surrey Audio-Visual Expressed Emotion (SAVEE) database. University of Surrey.

ISSN 2394 - 7780

[43] Amazon Web Services. (2022). Announcing model improvements and lower annotation limits for Amazon Comprehend custom entity recognition. (Context for SAVEE limitations via general dataset annotation issues)

[44] Dupuis, K., & Pichora-Fuller, M. K. (2010). Toronto Emotional Speech Set (TESS). University of Toronto, Psychology Department.

[45] Maulana, I. A. (2023). Replication Data for: Toronto emotional speech set (TESS). Telkom University Dataverse.

[46] Hou, X., et al. (2023). IMEMD-CRNN: An improved EMD-based convolutional recurrent neural network for speech emotion recognition. Frontiers in Psychology, 13, 1075624.

[47] ResearchGate. (n.d.). CREMA-D: Crowd-sourced emotional multimodal actor's dataset. (Context for TESS limitations via comparison)

[48] Amazon Web Services. (n.d.). Amazon Comprehend announces lower annotation limits for custom entity recognition. (Context for TESS limitations via general dataset annotation issues)

[49] ResearchGate. (n.d.). Description of TESS speech dataset. (Figure page, context for TESS use and limitations).

[50] Cao, H., et al. (2014). CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset. IEEE Transactions on Affective Computing, 5(4), 313-325.

[51] ResearchGate. (n.d.). CREMA-D: Crowd-sourced emotional multimodal actor's dataset.

[52] Zhang, Z., et al. (2023). Acoustic Knowledge-guided Transfer Learning with Dual Sparsity Constraint for Cross-Corpus Speech Emotion Recognition. arXiv preprint arXiv:2312.06466v1.

[53] Li, J., et al. (2023). Cross-Corpus Speech Emotion Recognition Based on Multi-Task Learning and Subdomain Adaptation. Entropy, 25(1), 124.

[54] Li, J., et al. (2023). Cross-Corpus Speech Emotion Recognition Based on Multi-Task Learning and Subdomain Adaptation. Entropy, 25(1), 124. [53]

[55] Xi, Y., et al. (2019). Speaker to Emotion: Domain Adaptation for Speech Emotion Recognition with Residual Adapters. APSIPA Annual Summit and Conference.

[56] TheMoonlight.io. (2024). Parameter-Efficient Finetuning for Speech Emotion Recognition and Domain Adaptation: A Review.