ISSN: 3078-5367 **DOI:** https://doi.org/10.69513/jncs.v2.i1.a5

Al-Noor Journal for Information Technology and Cybersecurity

https://jncs.alnoor.edu.iq/

A Comprehensive Review of Speech Emotion Recognition: Advances, Challenges, and Future Directions

[1] M K Hussein, [2]A Alqassab, [3]L T ALkahla, [4]D A Aliyu,

[1]Computer Center, Presidency of the University, University of Telafer, Mosul, Iraq.
[2]Software Department, Information Technology College, Ninevah University, Mosul, Iraq.
[3]Department of Computer Science, College of Education for Pure Science, University of Mosul, Mosul, Iraq .
[4]Department of Computing, Universiti Teknologi PETRONAS, Seri Iskandar, Perak 32610, Malaysia

Article information

Article history:
Received April, 15,2025
Revised May, 2,2025
Accepted June 20, 2025

Keywords:
Speech Emotion Recognition
Deep Learning, Transformer Models
Feature Extraction
Human-Computer Interaction
Multimodal Learning

Correspondence:

Lubna Thanoon ALkahla
lubna.thanoon@uoninevah.edu.iq

Abstract

Automated detection of human emotion from speech signals is a relatively new area in artificial intelligence aimed at determining the emotions people express through their speech. Traditionally, SER did feature extraction recognition with handcrafted ones and classical machine learning ones such as SVM (support vector machines) and HMM (hidden Markov models). The richness of emotions made these methodologies however challenging. The evolution of deep learning, in particular CNNs, RNNs, and other Transformer-based structures, has greatly improved the accuracy and robustness of SER systems. In this work, the SER is studied in depth taking into account the most relevant methods and feature extraction methods as well as an introduction of benchmark databases. It also includes augmentation methods, evaluation measures and the difficulties of real-time processing. Regardless of the advancements, SER continues to encounter challenges, including scarcity of datasets, imbalance between classes, domain adaptation, and high computational requirements. The review highlights unanswered questions regarding research and analyses. future directions, including multimodal fusion, self-supervised learning, and Explainable AI.

DOI: https://doi.org/10.69513/jncs.v2.i1.a5

Introduction

Speech Emotion Recognition (SER) deals with perceiving emotion in spoken language. As a subfield of artificial intelligence, it focuses on extracting various emotional states from human speech. The ability to machine to understand emotions impacts the domain of human-computer interaction (HCI), virtual assistants, healthcare, and even security applications [1]. Emotions play an important part in human communication where decisions, social affairs and a person's mental health are taken into account. In order to enhance user experience in applications such as call centers, mental health monitoring, SER systems analyze the acoustic, prosodic and linguistic features of speech to be able to categorize the emotion expressed [2].

Traditional techniques in SER used handcrafted features and, therefore, heavily depended upon the classical ML frameworks such as SVM, GMM, and HMM. These models also suffered from the restrictions of the poor generalization capacity and the poor characterization ability of complex emotional variations. The coming of age of deep learning techniques also was not an exception to the improvements seen in the case of SER as it did with the introduction of Convolutional Neural Networks

ISSN: 3078-5367 **DOI:** https://doi.org/10.69513/jncs.v2.i1.a5

(CNN), Recurrent Neural Networks (RNN) and Long-Short-Term-Memory units (LSTM) The classification aspect took a leap to a completely new level with these frameworks. More recent transformer-based architectures like Swin-Transformers and self-attention models have achieved improvements in the performance of SER by modelling long-range dependencies of signal [3][4].

Even with these advancements, some issues still persist in the SER field. There are still problems with an insufficient number of labeled emotional speech datasets, the diversity of emotions and how they are expressed across different cultures and languages, background noise, and still others. Real-time applications of speech emotion recognition systems also demand low power and latency requirements. Meeting these issues will require the aid of novel methods of feature extraction, data augmentation, and construction of deep learning architectures designed to perform in a noisy, real-world environment [4][5].

The purpose of this review is to analyze and summarize recent advances in speech emotion recognition systems, including feature extraction, speech signal processing, identification of deep learning networks, benchmark datasets, and evaluation metrics. We implemented a comparative analysis of approaches, documented important challenges, and explored newly emerged multimodal fusion, self-supervised learning, and real-time applications of speech emotion recognition systems. This review intends to explain to researchers the present condition of speech emotion recognition systems and the options available for subsequent investigations.

Datasets Overview

Datasets are essential for model construction in any SER system, and publicly available datasets are meant as a baseline for evaluation. These datasets are not uniform with respect to the number of speakers, emotions, recording environments, and whether they contain audio-visual components. Precise datasets amplify the effectiveness of models in real-life situations. On the contrary, most SER datasets suffer from class imbalance, which is the overrepresentation of some emotions and underrepresentation of others. There is ongoing research to address these limitations using data augmentation and transfer learning stratagies.

The Interactive Emotion Dyadic Motion Capture (IEMOCAP) acts as a benchmark in the field of SER. It contains texts of scripted and impromptu conversations from ten professional actors containing anger, happiness, sadness, and neutral dialogues. The dual audio and video dataset enables more advanced Multi-Modal SER research, broadening the scope of the IEMOCAP applications. Most researchers apply it to train their systems in speaker-independent mode, so the subjects of the training sessions differ from those who are tested [6].

Another dataset frequently referenced is the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). The dataset includes both spoken and sung expressions of emotion (24 actors portraying calm, happiness, sadness, anger, fear, surprise, and disgust), each emotion performed with varying intensity. This dataset is especially useful when examining the differences in emotional intensity. RAVDESS is commonly applied in machine learning contexts, particularly for systems using spectrogram representation as features [7].

Emo-DB, or the Berlin Database of Emotional Speech, is a German dataset that consists of recordings from ten actors performing seven emotions: anger, boredom, disgust, fear, happy, neutral, and sad. It is a well-balanced benchmark dataset of German emotional speech widely used in SER research [8]. The Toronto Emotional Speech Set (TESS) is another one which contains recordings of two female voice actors articulating target words with varying emotional inflections. TESS may be smaller, but it serves an important role in researching the effects of prosody and phonetic variation on emotion recognition [9].

Table 1: Datasets and Evaluation Metrics

Dataset	Language	Speakers	Emotions	Modality	Size (Samples)	Evaluation Metrics
IEMOCAP	English	10 (5M, 5F)	Angry, Happy, Sad, Neutral	Audio-Visual	5,531	WA, UA, F1-score
RAVDESS	English	24 (12M, 12F)	Calm, Happy, Sad, Angry, Fearful, Surprise, Disgust	Audio-Visual	2,452	Accuracy, WA, UA
Emo-DB	German	10 (5M, 5F)	Angry, Boredom, Disgust, Fear, Happy, Neutral, Sad	Audio	535	Accuracy, WA, UA
TESS	English	2 (F)	Angry, Happy, Neutral, Sad, Disgust, Fear, Surprise	Audio	2,800	Accuracy
SAVEE	English	4M	Angry, Happy, Neutral, Sad, Surprise, Fear, Disgust	Audio	480	Accuracy

WA (Weighted Accuracy): Class contributions are weighted according to size to correct class imbalance.
UA (Unweighted Accuracy): The average accuracy over all the classes.

F1-score (Precision and Recall): Harmonic mean of precision and recall, applicable for imbalanced datasets.

This table lists the characteristics of popular SER datasets in terms of language, speakers, emotional categories and evaluation protocol.

Methodologies for Speech Emotion Recognition (SER)

Speech Emotion Recognition (SER) is the task of identifying the emotions of a person from the speech signal and it involves looking at the computational methods by which emotional states are extracted from speech signals. SER approaches used: feature extraction, model architectures, data augmentation, and evaluation measures. In the last few years,

ISSN: 3078-5367 **DOI:** https://doi.org/10.69513/jncs.v2.i1.a5

advances in deep learning have considerably improved SER performances by automating feature extractions and leveraging much larger-scale datasets [10].

Feature Extraction Techniques

Feature extraction is the most important step in SER, since raw speech signals consist of both relevant and irrelevant information. Powerful features lead to the better performance of the model which emphasize the emotion-specific properties. Such features can be further grouped into prosodic features, spectral features, temporal features, and deep-learned features.

3.1.1 Prosodic Features

Prosodic features capture variations in pitch, energy, and speech duration, which convey emotional cues. These include [11]:

Fundamental Frequency (F0): Measures voice pitch, useful for detecting excitement or sadness.

Energy (Intensity): Indicates vocal effort, helping distinguish between calm and angry speech.

Speech Rate: Faster speech is often associated with happiness, while slower speech is linked to sadness.

3.1.2 Spectral Features

Spectral features analyze frequency distributions of speech signals and are commonly extracted using the Short-Time Fourier Transform (STFT) or Wavelet Transforms. Key spectral features include[12][13]:

Mel-Frequency Cepstral Coefficients (MFCCs): Most widely used features, capturing speech energy in different frequency bands.

Log-Mel Spectrograms: Represent the power of speech frequencies on a logarithmic scale.

Chroma Features: Capture harmonic content by analyzing pitch variations.

3.1.3. Temporal Features

Temporal features capture dynamic variations in speech signals over time. These include[14]:

Delta and Delta-Delta Features: First and second derivatives of MFCCs, capturing changes in speech patterns.

Zero Crossing Rate (ZCR): Counts the number of times the signal crosses zero amplitude, useful for detecting high-energy emotions.

3.1.4. Deep-Learned Features

Deep learning models, particularly Convolutional Neural Networks (CNNs) and Transformers, can learn high-level representations directly from spectrograms. These models extract hierarchical features that are more robust to noise and variations in speech[15][16].

Model Architectures for SER

Deep learning has revolutionized SER by eliminating the need for manual feature engineering. The primary architectures used in SER include CNN-based models, Recurrent Neural Networks (RNNs), hybrid CNN-LSTM models, and Transformer-based architectures.

3.2.1. Convolutional Neural Networks (CNNs)

CNNs are widely used in SER due to their ability to capture local spatial patterns in spectrograms. By applying convolutional filters, CNNs can learn emotion-specific features such as frequency modulation and formant structures[17][18].

Strengths: Robust to noise, efficient feature extraction.

Limitations: Limited capability in modeling long-term dependencies in speech.

Figure 1 illustrates a typical CNN architecture for SER, where speech signals are first converted into spectrograms before being processed by convolutional layers.

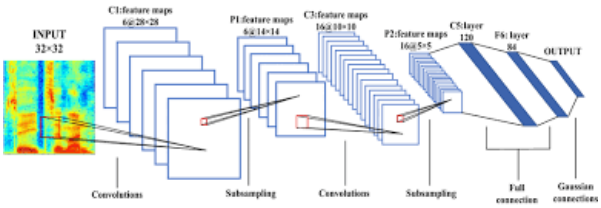

Figure 1: A CNN-based SER architecture using spectrogram inputs [18].

3.2.2. Recurrent Neural Networks (RNNs) and LSTMs

RNNs, particularly Long Short-Term Memory (LSTM) networks, are designed to capture temporal dependencies in sequential data. Since speech is inherently sequential, LSTMs are well-suited for SER [18].

Strengths: Effective in modeling speech dynamics and context.

Limitations: Computationally expensive, prone to vanishing gradient problems.

3.2.3. Hybrid CNN-LSTM Models

Hybrid models combine CNNs for feature extraction with LSTMs for temporal modeling. This approach leverages CNNs' ability to capture spatial patterns and LSTMs' capacity to model sequential dependencies [17][18].

Example: CNN extracts spectrogram features, which are then fed into an LSTM network for emotion classification.

Applications: Real-time emotion recognition, where CNN processes frames and LSTM integrates temporal dependencies.

3.2.4. Transformer-Based Models

Transformers, such as the Swin-Transformer, have gained popularity in SER due to their superior performance in capturing long-range dependencies. Unlike CNNs and RNNs, Transformers use self-attention mechanisms to process speech features in parallel, improving efficiency and accuracy [19].

Example: A Swin-Transformer model processes spectrogram inputs using shifted windows for better feature extraction.

ISSN: 3078-5367 **DOI:** https://doi.org/10.69513/jncs.v2.i1.a5

Benefits: Reduced computational complexity, ability to learn global speech features.

Data Augmentation and Preprocessing

To improve model robustness, data augmentation techniques are applied to increase dataset diversity. Common methods include [20][21]:

Noise Injection: Adding background noise to simulate real-world environments.

Pitch Shifting: Altering pitch levels to train models on speaker variations.

Time Stretching: Speeding up or slowing down speech without affecting pitch.

Spectrogram Masking: Randomly masking parts of spectrograms to prevent overfitting.

Table 2: Common Data Augmentation Techniques in SER

Technique	Description	Impact
Noise Injection	Adds background noise to speech signals	Improves robustness to real-world conditions
Pitch Shifting	Alters the pitch of the speech	Helps models generalize across different speakers
Time Stretching	Speeds up/slows down speech without affecting pitch	Improves model invariance to speech rate
Spectrogram Masking	Randomly removes parts of spectrogram	Prevents overfitting and improves generalization

Table 3: comparison between previous works

Reference	Year	Features Used	Model/Methodology	Datasets	Performance (WA, UA)	Key Contributions
Makhmudov et al. [22]	2022	MFCC, Spectrogram, Paralinguistic Features	Attention-Oriented Parallel CNN Encoders	EMO-DB, IEMOCAP	EMO-DB: WA 71.8%, UA 70.9%; IEMOCAP: WA 72.4%, UA 71.1%	Parallel CNN encoders with attention mechanism for feature fusion
Toyoshima et al.【23】	2023	Mel Spectrogram, GeMAPS	Multi-Input Deep Neural Network (CNN + DNN)	IEMOCAP	WA 66.57%, UA 61.49%	Focal loss for imbalanced data, GeMAPS integration for improved accuracy
Liu et al.【24】	2022	Log-Mels, Deltas, Delta-Deltas	Multitask Learning with Cascaded Attention and Self-Adaptive Loss	IEMOCAP	WA 80.47%, UA 77.56%	Non-personalized features, cascaded attention network, multitask learning with gender and continuous attributes
Aghajani et al.【25】	2020	Scalogram-based Features	CNN + RNN with Attention	RAVDESS, SAVEE, Emo-DB	SAVEE: 83.9% (UAR), Emo-DB: 86.4% (UAR)	3D Scalogram as input, CNN for local features, RNN for long-term dependencies
Liao et al.[26】	2023	Log Spectrogram, Log-Mel Spectrogram	Swin-Transformer	IEMOCAP	WA 70.1%, UA 62.4%	Application of Swin-Transformer for SER, optimized spectrogram selection
Mustaqeem & Kwon [27]	2020	Raw Speech Data, ConvLSTM Features	Hierarchical ConvLSTM + Sequence Learning	IEMOCAP, RAVDESS	IEMOCAP: 75%, RAVDESS: 80%	Novel hierarchical ConvLSTM model with adaptive GRU-based sequence learning and center loss.
Pan & Wu [28]	2023	MFCC, Mel Spectrogram	1D CNN + LSTM + Data Augmentation	RAVDESS, Emo-DB, IEMOCAP	RAVDESS: 90.6%, Emo-DB: 96.7%	Combined CNN and LSTM with data augmentation to improve accuracy and generalization.
Barhoumi & Ben Ayed [29]	2023	MFCC, Zero Crossing Rate, Chroma Features	CNN + BiLSTM with Data Augmentation	TESS, EmoDB, RAVDESS	RAVDESS: 85%, EmoDB: 90%	Real-time SER system integrating deep learning models

Discussion of Results

The advancements in Speech Emotion Recognition (SER) have been driven primarily by deep learning models, which have outperformed traditional machine learning approaches by automating feature extraction and capturing complex emotional patterns. This section discusses key findings from the reviewed methodologies, evaluates their performance, and identifies critical challenges and future opportunities.

1. Performance of Deep Learning Models

CNN-based models excel in extracting spatial features from spectrograms, making them effective for emotion recognition tasks. However, their inability to model long-term temporal dependencies limits their performance in dynamic speech signals.

RNNs and LSTMs address this limitation by capturing sequential dependencies, improving accuracy in emotion classification. However, they suffer from high computational costs and vanishing gradient problems.

Hybrid CNN-LSTM models combine the strengths of both architectures, achieving better performance by leveraging CNNs for feature extraction and LSTMs for temporal modeling. These models have shown success in real-time SER applications.

Transformer-based models (e.g., Swin-Transformers) have emerged as a powerful alternative, utilizing self-attention mechanisms to process long-range dependencies efficiently. They outperform traditional models in accuracy but require large datasets and significant computational resources.

2. Impact of Feature Extraction Techniques

Prosodic features (e.g., pitch, energy, speech rate) remain crucial for distinguishing emotions like anger (high energy) and sadness (low pitch).

Spectral features (e.g., MFCCs, log-Mel spectrograms) provide robust representations of speech signals, improving model generalization.

Deep-learned features (automatically extracted by CNNs or Transformers) reduce reliance on manual feature engineering and enhance adaptability to diverse datasets.

3. Challenges in SER

Dataset Limitations: Most SER datasets (e.g., IEMOCAP, RAVDESS) suffer from class imbalance, limited speaker diversity, and cultural bias. Data augmentation (e.g., noise injection, pitch shifting) helps mitigate these issues but does not fully address the lack of large-scale, annotated datasets.

Real-Time Processing: Deploying SER systems in real-world applications (e.g., call centers, virtual assistants) requires low-latency models, which remains a challenge for complex architectures like Transformers.

Cross-Lingual and Cross-Cultural Generalization: Current models perform well on English and German datasets but struggle with

ISSN: 3078-5367 **DOI:** https://doi.org/10.69513/jncs.v2.i1.a5

underrepresented languages and cultural variations in emotional expression.

Explainability: Deep learning models, particularly Transformers, are often treated as "black boxes." Incorporating Explainable AI (XAI) techniques is essential for building trust in SER applications, especially in healthcare and security.

4. Future Directions

Multimodal Fusion: Combining speech with facial expressions, text, or physiological signals (e.g., EEG) can improve emotion recognition accuracy.

Self-Supervised Learning: Leveraging unlabeled speech data through contrastive learning or autoencoders can reduce dependency on annotated datasets.

Lightweight Architectures: Developing efficient models (e.g., knowledge distillation, quantization) for edge devices will enable real-time SER deployment.

Cross-Domain Adaptation: Transfer learning and domain adaptation techniques can enhance model performance across different languages and recording conditions.

Conclusion

Speech Emotion Recognition (SER) has significantly advanced, largely thanks to deep learning models like CNNs, RNNs, and Transformers replacing older methods. By automatically detecting and learning relevant features and patterns over time, these new methods are particularly good at extracting the complex emotional subtleties embedded in speech. Although benchmark datasets like IEMOCAP and RAVDESS have facilitated advancement by offering a shared foundation for system comparisons, the community recognizes their shortcomings and acknowledges the perpetual need for richer and more diverse data to surpass existing boundaries.

Even with the advancements, the primary issues related to the presence of rich, diverse data, different languages and cultures, and the practicality of implementing SER systems in real-time applications are still unsolved. Addressing these problems anticipates future work: improving and enriching the dataset by augmentation methods, speech with video or

References:

1. N. Naeeni and B. Nasersharif, "Feature and classifier-level domain adaptation in DistilHuBERT for cross-corpus speech emotion recognition," *Computers in Biology and Medicine*, vol. 194, p. 110510, Jun. 2025, doi: 10.1016/j.compbiomed.2025.110510.
2. Z. Shah, S. Zhiyong, and Adnan, "Enhancements in immediate speech emotion detection: harnessing prosodic and spectral characteristics," *International Journal of Innovative Science and Research Technology (IJISRT)*, pp. 1526–1534, May 2024, doi: 10.38124/ijisrt/ijisrt24apr872.
3. J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, Sep. 2018, doi: 10.1016/j.bspc.2018.08.035.
4. W. Zhao and Z. Yang, "An emotion speech synthesis method based on VITS," *Applied Sciences*, vol. 13, no. 4, p. 2225, Feb. 2023, doi: 10.3390/app13042225.
5. Z. Shah, S. Zhiyong, and Adnan, "Enhancements in immediate speech emotion detection: harnessing prosodic and spectral characteristics," *International Journal of Innovative Science and Research Technology (IJISRT)*, pp. 1526–1534, May 2024, doi: 10.38124/ijisrt/ijisrt24apr872.
6. C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008, doi: 10.1007/s10579-008-9076-6.
7. S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLoS ONE*, vol. 13, no. 5, p. e0196391, 2018, doi: 10.1371/journal.pone.0196391.
8. F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Interspeech*, Lisbon, Portugal, 2005, pp. 1517–1520.
9. K. R. Scherer, R. Banse, and H. G. Wallbott, "Emotion inferences from vocal expression correlate across languages and cultures," *J. Cross-Cult. Psychol.*, vol. 32, no. 1, pp. 76–92, 2001
10. D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 59, p. 101894, Feb. 2020, doi: 10.1016/j.bspc.2020.101894.
11. F. Daneshfar, S. J. Kabudian, and A. Neekabadi, "Speech emotion recognition using hybrid spectral-prosodic features of speech signal/glottal waveform, metaheuristic-based dimensionality reduction, and Gaussian elliptical basis function network classifier," *Applied Acoustics*, vol. 166, p. 107360, Apr. 2020, doi: 10.1016/j.apacoust.2020.107360.
12. L. Abdel-Hamid, "Egyptian Arabic speech emotion recognition using prosodic, spectral and wavelet features," *Speech Communication*, vol. 122, pp. 19–30, May 2020, doi: 10.1016/j.specom.2020.04.005.
13. Z. Shah, S. Zhiyong, and Adnan, "Enhancements in immediate speech emotion detection: harnessing prosodic and spectral characteristics," *International Journal of Innovative Science and Research Technology (IJISRT)*, pp. 1526–1534, May 2024, doi: 10.38124/ijisrt/ijisrt24apr872.
14. M. R. Schädler, B. T. Meyer, and B. Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 131, no. 5, pp. 4134–4151, May 2012, doi: 10.1121/1.3699200.
15. D. Wang and J. Chen, "Supervised Speech Separation Based on Deep Learning: An Overview," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 26, no. 10, pp. 1702–1726, May 2018, doi: 10.1109/taslp.2018.2842159.
16. N. Mustaqeem and S. Kwon, "A CNN-Assisted enhanced audio signal processing for speech emotion recognition," *Sensors*, vol. 20, no. 1, p. 183, Dec. 2019, doi: 10.3390/s20010183.
17. J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, Sep. 2018, doi: 10.1016/j.bspc.2018.08.035.
18. T. Anvarjon, N. Mustaqeem, and S. Kwon, "Deep-Net: a lightweight CNN-Based speech emotion recognition system using deep frequency features," *Sensors*, vol. 20, no. 18, p. 5212, Sep. 2020, doi: 10.3390/s20185212.
19. Z. Liao and S. Shen, "Speech emotion recognition based on Swin-Transformer," *Journal of Physics Conference Series*, vol. 2508, no. 1, p. 012056, May 2023, doi: 10.1088/1742-6596/2508/1/012056.
20. B. T. Atmaja and A. Sasou, "Effects of data augmentations on speech emotion recognition," *Sensors*, vol. 22, no. 16, p. 5941, Aug. 2022, doi: 10.3390/s22165941.

ISSN: 3078-5367 **DOI:** https://doi.org/10.69513/jncs.v2.i1.a5

21. D. S. Park *et al.*, "SpecAugment: a simple data augmentation method for automatic speech recognition," *Interspeech 2022*, Sep. 2019, doi: 10.21437/interspeech.2019-2680.
22. F. Makhmudov, A. Kutlimuratov, F. Akhmedov, M. S. Abdallah, and Y.-I. Cho, "Modeling speech Emotion Recognition via Attention-Oriented Parallel CNN encoders," *Electronics*, vol. 11, no. 23, p. 4047, Dec. 2022, doi: 10.3390/electronics11234047.
23. I. Toyoshima, Y. Okada, M. Ishimaru, R. Uchiyama, and M. Tada, "Multi-Input Speech Emotion Recognition Model using MEL Spectrogram and GEMAPS," *Sensors*, vol. 23, no. 3, p. 1743, Feb. 2023, doi: 10.3390/s23031743.
24. Y. Liu *et al.*, "A multitask learning approach based on cascaded attention network and Self-Adaption loss for speech emotion recognition," *IEICE Transactions on Fundamentals of Electronics Communications and Computer Sciences*, vol. E106.A, no. 6, pp. 876–885, Dec. 2022, doi: 10.1587/transfun.2022eap1091.
25. K. Aghajani and E. P. Afrakoti I., "Speech emotion recognition using Scalogram based deep structure," *International Journal of Engineering. Transactions B: Applications*, vol. 33, no. 2, Feb. 2020, doi: 10.5829/ije.2020.33.02b.13.
26. Z. Liao and S. Shen, "Speech emotion recognition based on Swin-Transformer," *Journal of Physics Conference Series*, vol. 2508, no. 1, p. 012056, May 2023, doi: 10.1088/1742-6596/2508/1/012056.
27. N. Mustaqeem and S. Kwon, "CLSTM: Deep Feature-Based Speech Emotion Recognition using the Hierarchical CONVLSTM network," *Mathematics*, vol. 8, no. 12, p. 2133, Nov. 2020, doi: 10.3390/math8122133.
28. S.-T. Pan and H.-J. Wu, "Performance Improvement of Speech Emotion Recognition Systems by Combining 1D CNN and LSTM with Data Augmentation," *Electronics*, vol. 12, no. 11, p. 2436, May 2023, doi: 10.3390/electronics12112436.
29. C. Barhoumi and Y. BenAyed, "Real-time speech emotion recognition using deep learning and data augmentation," *Artificial Intelligence Review*, vol. 58, no. 2, Dec. 2024, doi: 10.1007/s10462-024-11065-x.