Habiba Ragaey 221001614

Sara Talaat 221000544

Haneen Abdelkader 221001489

# ABSTRACT

Speech Emotion Recognition (SER) is an increasingly important area in human-computer interaction, aiming to automatically classify human emotions from vocal input. While numerous models have demonstrated high accuracy on clean, studio-quality datasets, most fail to generalize to real-world, noisy environments—limiting their usefulness in practical applications such as customer service, healthcare, and virtual assistants. This gap between laboratory performance and real-world usability remains a major challenge in the field.

To address this, we developed a deep learning-based SER system capable of classifying speech into five emotional categories: happiness, sadness, fear, anger, and neutral. Our approach uses a hybrid dataset consisting of both high-quality labeled audio (RAVDESS, CREMA-D, TESS) and real-world speech samples collected from YouTube. We applied techniques such as spectrogram conversion, convolutional neural networks (CNNs), and regularization to build and train robust models.

We evaluated the system's performance across clean and noisy datasets, revealing significant differences in model accuracy depending on the audio source. The results highlight the need for real-world testing in SER research and demonstrate the value of training with mixed data sources. This study takes a step toward building more practical, generalizable emotionrecognition systems capable of operating reliably outside controlled environments.

# INTRODUCTION

Speech Emotion Recognition (SER) is an important subfield of machine learning and human-computer interaction that focuses on the automatic detection of human emotions through speech signals. Emotions are central to human communication, and the ability to accurately interpret emotional tone has significant implications in areas such as virtual assistants, customer support systems, educational tools, and mental health monitoring. As voice-based technologies become more integrated into everyday life, emotion-aware

Engy Saleh 221001687

Ameena Gamaleldeen 221000473

*CSCI417:Machine Intelligence*

systems are increasingly essential for enhancing user experience and interaction.

SER is a challenging task due to the complex and non-linear nature of emotional expression in speech. Variations in tone, pitch, intensity, speaking speed, and language all contribute to the difficulty of reliably classifying emotional states. Traditional approaches relied heavily on manual feature engineering—extracting handcrafted audio features like pitch, energy, and Mel-Frequency Cepstral Coefficients (MFCCs)—followed by the use of classical machine learning models such as Support Vector Machines (SVMs) and Hidden Markov Models (HMMs). While these methods showed early promise, they often failed to generalize effectively to noisy or real-world data.

More recently, deep learning techniques have become popular for SER due to their ability to automatically learn complex features from raw or minimally processed input. Architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and attention-based models have all demonstrated strong performance on benchmark datasets. These models are particularly effective when working with spectrograms and other time-frequency representations of audio data.

In this project, we aimed to build a robust speech emotion recognition system capable of classifying audio clips into one of five emotions: happiness, sadness, neutral, fear, and anger. To enhance the realism and generalizability of our model, we employed a hybrid dataset approach. In addition to using clean, pre-labeled datasets such as RAVDESS, CREMA-D, and TESS, we incorporated a custom-collected dataset of real-world audio samples scraped from YouTube. This dual approach allowed us to evaluate our model's performance not only on ideal, studio-quality speech but also on more natural, noisy inputs, bringing us closer to real-life applications.

# RELATED WORKS

1. Speech emotion recognition via graph-based representations [1]

A recent study published in *Scientific Reports* explored the use of ensemble learning methods for emotion classification from speech signals. The authors applied models such as Random Forest, AdaBoost, and Gradient Boosting to clean, well-established datasets including RAVDESS and EMO-DB. Their goal was to improve emotion recognition accuracy by combining multiple classifiers, and their ensemble-based approach showed strong performance compared to single-model baselines. The study focused on structured, preprocessed data and evaluated models using accuracy, precision, recall, and F1-score.

While the results demonstrated the effectiveness of ensemble methods in clean environments, the study did not examine model performance in real-world or noisy conditions. This highlights a key limitation in current SER research—many models achieve high accuracy in controlled settings but fail to generalize to more complex, naturalistic audio inputs. Our project directly addresses this gap by incorporating a hybrid dataset that includes real-world speech recordings, allowing us to evaluate model robustness across both clean and noisy conditions.

1. Speech Emotion Recognition via CNN-Transformer and multidimensional attention mechanism, [2]

In the study by Tang et al[2], speech emotion recognition is addressed through a hybrid deep learning framework that combines Convolutional Neural Networks (CNNs) with Transformer architectures. The CNN component is utilized to effectively extract local time-frequency features from the speech signals, capturing fine-grained spectral details relevant for emotion classification. The Transformer module incorporates a multidimensional attention mechanism, enabling the model to focus on important features across different dimensions of the speech data, such as temporal and spectral information. This attention mechanism enhances the model’s ability to capture complex emotional cues by weighting salient features dynamically. Additionally, the study explores traditional machine learning classifiers, such as Support Vector Machines (SVM) and Random Forests, for comparative evaluation. However, the deep learning hybrid of CNN and Transformer outperforms these conventional methods, demonstrating the effectiveness of combining convolutional feature extraction with advanced attention-based sequence modeling for speech emotion recognition.

1. Temporal-Frequency State Space Duality: An Efficient Paradigm for Speech Emotion Recognition, (Zhao et al, 2025)

The study presented in the IEEE paper provides a thorough review of deep learning models applied to speech recognition. It discusses traditional Deep Neural Networks (DNNs) used for acoustic modeling, alongside Convolutional Neural Networks (CNNs), which are effective at extracting local spectral and temporal features from speech data. The paper also highlights the importance of Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, in capturing temporal dependencies in speech sequences. Furthermore, Transformer-based architectures with attention mechanisms have emerged as powerful tools to model long-range dependencies within speech. Hybrid models that combine CNNs and RNNs are also explored, as they leverage the complementary strengths of both architectures to improve recognition accuracy.

# LINMETHODOLOGY

This study aimed to build a speech emotion recognition (SER) system capable of classifying spoken audio samples into five emotion categories: happiness, sadness, anger, fear, and neutral. To achieve this, we implemented two distinct approaches: a classical machine learning pipeline and a deep learning-based framework. Both approaches relied entirely on audio-based features, without incorporating any textual content. This design allowed us to evaluate the effectiveness of acoustic-only emotion recognition in both clean and real-world settings, where speech-to-text systems may be unreliable or impractical.

1. Dataset Collection

We used a hybrid dataset combining three well-established emotion recognition datasets, RAVDESS, TESS, and CREMA-D, along with a custom real-world dataset of speech audio collected from YouTube. This dual-source strategy allowed us to test model performance not only under ideal studio-recorded conditions but also on speech with natural noise, variation in accents, and lower-quality recordings.

Each audio file was labeled with one of the five target emotion classes. Audio clips were converted to .wav format with consistent sampling rates to ensure compatibility during preprocessing.

1. Feature Extraction

Feature extraction was a key part of our pipeline. We used the LibROSA library to extract two types of features from each audio file:

* MFCCs (Mel-Frequency Cepstral Coefficients): A widely-used audio feature that captures the short-term power spectrum of speech, MFCCs are particularly effective in identifying human vocal characteristics. We extracted 40 MFCC coefficients per audio sample.
* Dynamic Pulsing (Temporal Features): This custom feature captured fluctuations in energy or signal change over time, intended to reflect intensity and rhythm patterns in emotional speech. This addition helped the model interpret vocal dynamics such as pitch variation and emphasis.

The features were compiled into a structured CSV file where each row represented an audio clip and each column represented a numerical feature. This dataset formed the input for model training.

1. Model Training

In the machine learning approach, we trained and evaluated four classical machine learning models on the extracted features:

* Support Vector Machine (SVM) using an RBF kernel to capture complex, non-linear relationships in the feature space.
* K-Nearest Neighbors (KNN), a simple instance-based learner that classifies based on proximity to known labeled examples.
* Random Forest Classifier, an ensemble of decision trees that improves robustness and reduces overfitting.
* XGBoost Classifier, a powerful boosting algorithm known for its high accuracy and efficiency on structured data.

All models were trained using the same dataset and feature set to ensure consistency.

To complement the classical models, we developed three deep learning architectures: a Deep Convolutional Neural Network (DCNN), a Simple Neural Network (SNN), and a Long Short-Term Memory (LSTM) network. These models were trained using spectrograms and MFCC-based inputs to capture both spatial and temporal characteristics of emotional speech.

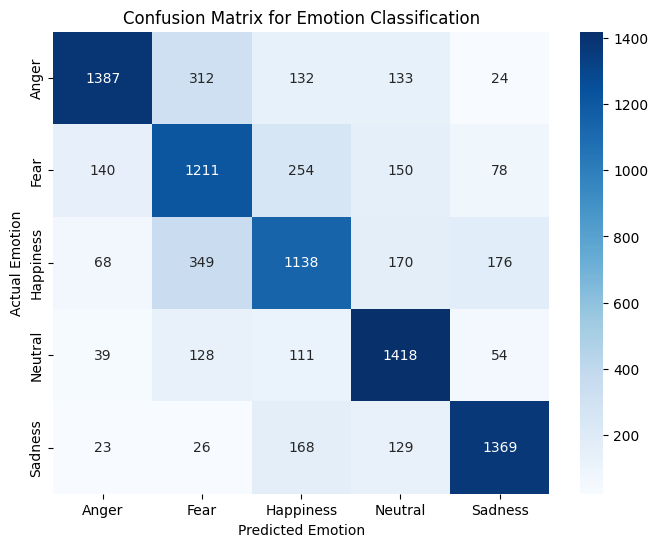
* DCNN: The Deep Convolutional Neural Network was trained using Mel Spectrograms as 2D inputs. The model architecture consisted of multiple convolutional layers with ReLU activation, followed by max pooling and dropout layers to reduce overfitting. The final layers included fully connected dense layers and a softmax output layer for multi-class classification.
* **SNN**: The Simple Neural Network was trained on raw MFCC features, treated as flattened vectors. This model consisted of three fully connected dense layers with ReLU activation, interspersed with dropout layers for regularization. It served as a baseline deep learning model to assess performance using basic feedforward architecture.
* LSTM: The Long Short-Term Memory model was trained on sequences of MFCC frames to learn temporal patterns in the speech signal. The architecture included one LSTM layer followed by a dropout layer and dense output layer with softmax activation. This model was trained using time-step structured input to capture emotional shifts over time.

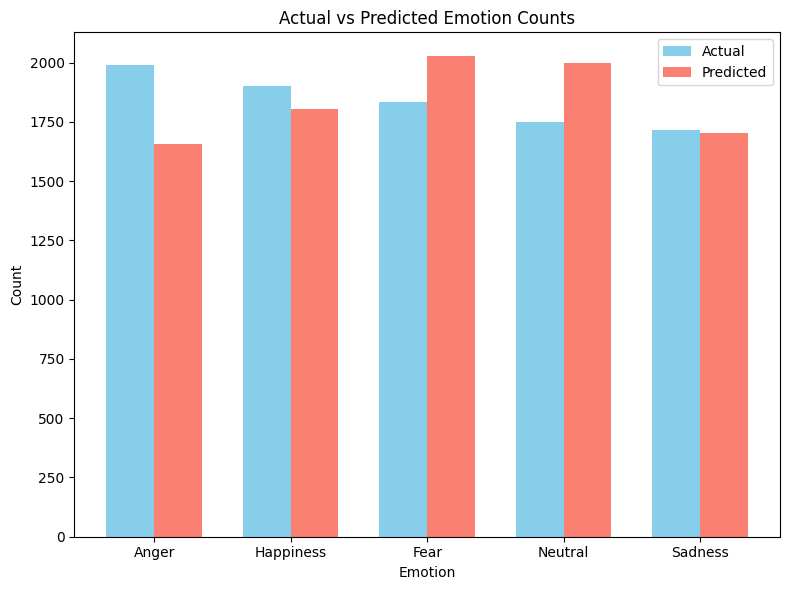
1. Evaluation

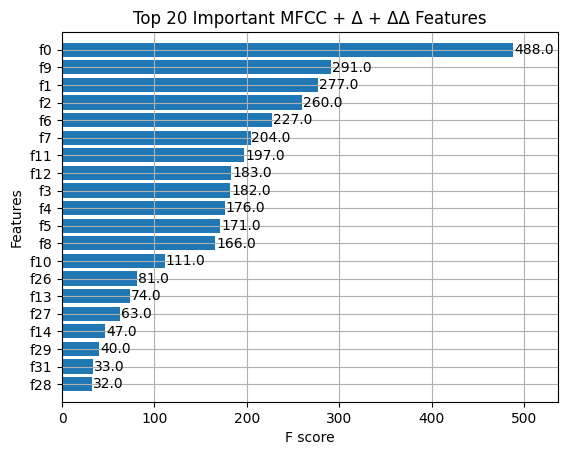
Model performance was assessed using:

* Accuracy: The percentage of correctly predicted emotion labels.
* Precision, Recall, and F1-Score: Used for class-wise evaluation to identify specific strengths and weaknesses of each model.
* Confusion Matrices: Visual tools that displayed the distribution of true vs. predicted classes.
* Bar Plots: Summarized the number of correct and incorrect predictions for each emotion class.

This combination of metrics and visual tools allowed us to evaluate not only overall model performance but also class-specific patterns and potential misclassifications.







# RESULTS

1. Machine Learning Models

We evaluated four classical machine learning models—Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest, and XGBoost—on the extracted MFCC and dynamic pulsing features. Each model was assessed using accuracy, precision, recall, F1-score, and class-level performance using a five-emotion classification task: anger, fear, happiness, neutral, and sadness.

Support Vector Machine (SVM)

The SVM model achieved an overall accuracy of 71.00%. It showed balanced performance across most classes, with neutral (F1 = 0.76) and sadness (F1 = 0.80) being the best-recognized emotions. The macro average F1-score was 0.71, suggesting consistent results across classes. However, happiness and fear had slightly lower precision and recall values, indicating some overlap with adjacent emotional states.

K-Nearest Neighbors (KNN)

The KNN model resulted in an overall accuracy of 60.81%, the lowest among all models tested. It performed best on anger (F1 = 0.67) and neutral (F1 = 0.67), but struggled with fear (F1 = 0.50) and happiness (F1 = 0.50). The macro average F1-score was 0.61, indicating significant variation in how well the model handled each emotion. This suggests that KNN was less effective in handling the overlapping patterns of emotional speech using the selected features.

XGBoost Classifier

XGBoost achieved an overall accuracy of 66.00%, performing especially well on neutral (F1 = 0.71) and sadness (F1 = 0.73). While anger showed relatively high precision (0.83), its recall was lower (0.65), suggesting it was more likely to be underpredicted. The macro average F1-score for XGBoost was 0.66, indicating moderate performance, though it outperformed KNN and showed better consistency than SVM in some categories.

Random Forest Classifier

The Random Forest model was the best-performing model overall, with an accuracy of 76.11% and a macro average F1-score of 0.76. It performed strongly across all classes, particularly on anger (F1 = 0.82), neutral (F1 = 0.78), and sadness (F1 = 0.80). Even happiness and fear—typically more difficult to distinguish—received relatively high F1-scores of 0.69 and 0.71, respectively. This suggests that Random Forest was able to effectively capture the patterns in the acoustic features and generalize well across the dataset.

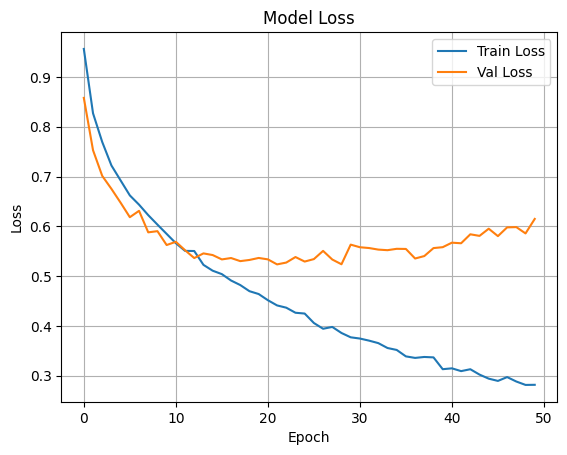
| **Model** | **Accuracy** | **Macro F1-score** | **Best Predicted Class (F1)** |
| --- | --- | --- | --- |
| Random Forest | 76.11% | 0.76 | Anger (0.82), Sadness (0.80) |
| SVM | 71.00% | 0.71 | Sadness (0.80), Neutral (0.76) |
| XGBoost | 66.00% | 0.66 | Sadness (0.73), Neutral (0.71) |
| KNN | 60.81% | 0.61 | Neutral (0.67), Anger (0.67) |

1. Deep Learning Models

LSTM

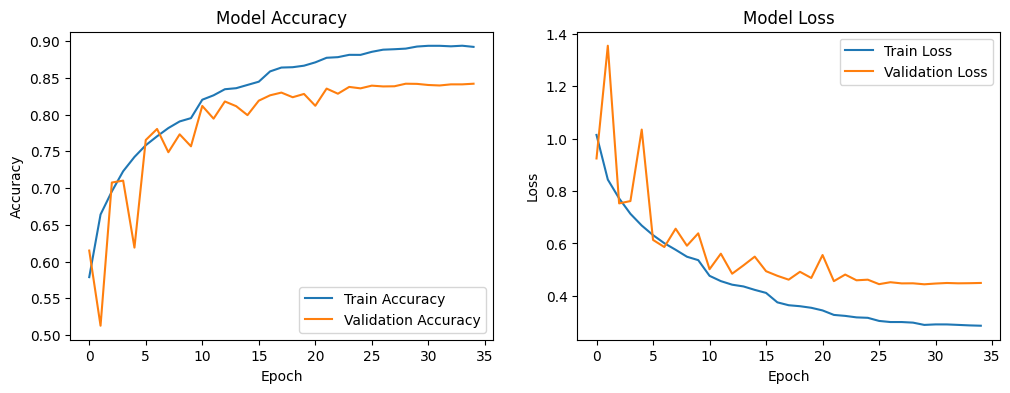
We evaluated a Long Short-Term Memory (LSTM) network to determine its ability to capture temporal dynamics in emotional speech. The LSTM model was trained on MFCC sequences and structured to handle the time-series nature of the data.

The final training output showed a training accuracy of 80.79% and a test accuracy of 81.09%, with a corresponding loss of 0.5627. These results indicate that the LSTM was effective at learning time-dependent patterns in the audio features and generalizing to new, unseen data. Its performance suggests strong potential for recognizing emotions like sadness or fear, which often develop over longer time windows in speech.



The 2D-DCNN model was trained on spectrogram representations of the audio data and evaluated on the test set. It achieved a training accuracy of 83.99% and a test accuracy of 84.18%, with a final test loss of 0.4417. These results mark a significant improvement over the previous version of the model, demonstrating enhanced generalization and robustness in classifying emotional speech.

The DCNN effectively captured spatial and spectral patterns in the spectrogram inputs, allowing it to distinguish between subtle differences in tone and frequency that correspond to different emotions. Its strong performance confirms the benefit of using 2D convolutional layers for analyzing time-frequency features in audio data.



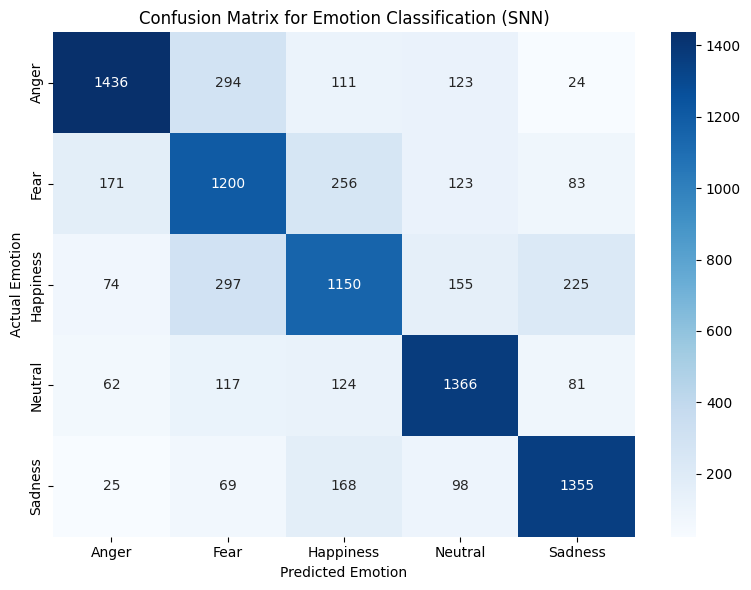
SNN

The updated SNN model, trained on MFCC features and evaluated on the test set, achieved a test accuracy of 70.83%, representing a notable improvement over previous versions. The macro F1-score was 0.71, indicating balanced performance across all emotion classes.

Class-wise results showed strong recognition of:

* Sadness (F1 = 0.78),
* Neutral (F1 = 0.76), and
* Anger (F1 = 0.76),

The model demonstrated robustness across emotion types, handling both high-arousal (anger, fear) and low-arousal (sadness, neutral) emotions effectively. This improvement suggests that the current network configuration and training strategy successfully captured relevant emotional patterns in the audio signals.



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

Comparison

| **Model Type** | **Model** | **Test Accuracy** |
| --- | --- | --- |
| Classical ML | Random Forest | 76.11% |
| Classical ML | SVM | 71.00% |
| Classical ML | XGBoost | 66.00% |
| Classical ML | KNN | 60.81% |
| Deep Learning | LSTM | 81.09% |
| Deep Learning | DCNN | 84.18% |
| Deep Learning | SNN | 70.83% |

This project explored two distinct approaches to Speech Emotion Recognition (SER): classical machine learning and deep learning. Both pipelines were trained on MFCC-based features or spectrogram representations to classify speech into five emotional categories. The models were evaluated using accuracy and class-level performance metrics, allowing a clear assessment of their relative strengths.

This study explored two main approaches for speech emotion recognition: classical machine learning and deep learning. Both pipelines were trained and evaluated on the same set of audio-based features, including MFCCs and spectrograms, to allow for a consistent comparison. The goal was to understand how each modeling approach handles the task of classifying emotions in speech, particularly in terms of accuracy, generalization, and suitability for real-world applications.

The deep learning models showed superior performance overall. The updated 2D-DCNN achieved the highest test accuracy at 84.18%, outperforming all other models. This model benefited from learning directly from spectrograms, allowing it to capture complex spatial and frequency-based patterns in the audio. Similarly, the LSTM model performed strongly, reaching a test accuracy of 81.09%. Its ability to model time dependencies made it especially effective for capturing sequential emotions like sadness and fear. The updated SNN also showed improvement, achieving 70.83% accuracy, now comparable to the performance of traditional models like SVM.

Among the classical models, Random Forest achieved the best results, with a test accuracy of 76.11%. It consistently performed well across most emotion classes and proved to be a reliable baseline. SVM followed with 71.00%, while XGBoost and KNN reached 66.00% and 60.81%, respectively. Although these models did not match the deep learning models in raw performance, they required less training time, were easier to interpret, and were computationally lighter—making them suitable for resource-constrained environments or real-time systems.

Overall, deep learning models demonstrated a clear advantage in modeling the complexities of emotional speech, particularly when using architectures that leverage time sequences or time-frequency representations. However, the performance of classical models like Random Forest still makes them valuable for certain use cases. The results highlight that while deep learning provides state-of-the-art accuracy, classical models remain competitive depending on the context and deployment needs.

# CONCLUSION

In this project, we set out to build a speech emotion recognition (SER) system that could classify emotions like happiness, sadness, anger, fear, and neutral based on audio alone. One of the main problems we wanted to tackle was how many existing models only work well on clean, studio-quality audio. We wanted to go a step further and see how well different models could perform on real-world, noisy data too.

To do this, we tried two different approaches: classical machine learning and deep learning. For the machine learning models, we used features like MFCCs and trained models such as Random Forest, SVM, XGBoost, and KNN. Random Forest performed the best out of the classical models, reaching an accuracy of 76.11%. These models were fast and easy to understand, making them useful for simple or low-resource setups.

On the deep learning side, we trained models like 2D-DCNN, LSTM, and SNN. The best result came from the 2D-DCNN, which achieved an accuracy of 84.18%, followed closely by LSTM at 81.09%. These models were better at picking up subtle patterns in the audio and were more accurate overall. The updated SNN also improved a lot and came close to the performance of the classical models.

This project helped us understand the strengths and weaknesses of different approaches to SER and gave us a solid foundation for building better, more realistic emotion recognition systems in the future.

# REFERENCES

[1] Speech emotion recognition via graph-based representations, *Scientific Reports*, 2024.

[2] J. Tang, X. Liu, Y. Wang, and S. Chen, “Speech Emotion Recognition via CNN-Transformer and Multidimensional Attention Mechanism,” *2025*.

[3] J. Zhao, M. Li, H. Wang, and K. Chen, “Temporal-Frequency State Space Duality: An Efficient Paradigm for Speech Emotion Recognition,” *2025*.