Journal of Computer and Communications, 2025, 13(3), 103-115
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Speech Emotion Recognition Based on CNN-Transformer with Different Loss Function

Bin Li

Department of Mental Health Education, Nanjing Audit University, Nanjing, China
Email: lb8976@126.com

How to cite this paper: Li, B. (2025) Speech Emotion Recognition Based on CNN-Transformer with Different Loss Function. *Journal of Computer and Communications*, **13**, 103-115.
https://doi.org/10.4236/jcc.2025.133008

Received: March 11, 2025
Accepted: March 23, 2025
Published: March 26, 2025

Abstract

Speech Emotion Recognition (SER) is crucial for enhancing human-computer interactions by enabling machines to understand and respond appropriately to human emotions. However, accurately recognizing emotions from speech is challenging due to variations across speakers, languages, and environmental contexts. This study introduces a novel SER framework that integrates Convolutional Neural Networks (CNNs) for effective local feature extraction from Mel-Spectrograms, and transformer networks employing multi-head self-attention mechanisms to capture long-range temporal dependencies in speech signals. Additionally, the paper investigates the impact of various loss functions—L1 Loss, Smooth L1 Loss, Binary Cross-Entropy Loss, and Cross-Entropy Loss—on the accuracy and generalization performance of the model. Experiments conducted on a combined dataset formed from RAVDESS, SAVEE, and TESS demonstrate that the CNN-Transformer model with Cross-Entropy Loss achieves superior accuracy, outperforming other configurations. These findings highlight the importance of appropriately selecting loss functions to enhance robustness and effectiveness in speech emotion recognition systems.

Keywords

Speech Emotion Recognition, CNN-Transformer, Mel-Spectrogram, Multi-Head Self-Attention, Loss Function

1. Introduction

Speech Emotion Recognition (SER) [1] [2] is a critical area in human-computer interaction, aiming to identify and classify human emotions from speech signals. Emotions play a vital role in communication, influencing how messages are perceived and understood. With the increasing integration of artificial intelligence in

daily life, from virtual assistants to customer service bots, the ability to accurately recognize emotions from speech has become essential for creating more natural and empathetic interactions. Traditional SER systems have relied on handcrafted features such as the MFCC [3] [4], harmonic to noise ratio (HNR) [5], spectrogram [6], mel-spectrogram [7], etc., to find SER. However, these methods often struggle with the variability and complexity of emotional expressions across different speakers, languages, and contexts. Recent advancements in deep learning have revolutionized SER by enabling the automatic extraction of relevant features from raw audio data. Among these, Convolutional Neural Networks (CNNs) [8]-[10] and Transformers [11] [12] have emerged as powerful tools for capturing both local and global patterns in speech signals.

CNNs are particularly effective at extracting hierarchical features from spectrograms or other time-frequency representations of speech. They excel at capturing local patterns, such as short-term variations in pitch and intensity, which are crucial for emotion recognition. On the other hand, Transformers, originally developed for natural language processing, have shown remarkable success in modeling long-range dependencies in sequential data [13]. By leveraging self-attention mechanisms [14], Transformers can capture the global context of speech signals, which is essential for understanding the overall emotional tone.

Despite the strengths of CNNs and Transformers, the choice of loss function [15] plays a pivotal role in the performance of SER systems. Loss functions guide the learning process by quantifying the difference between predicted and actual emotions [16]. Commonly used loss functions, such as Cross-Entropy Loss [17], focus on minimizing classification errors. However, emotions in speech are often subtle and can overlap, making it challenging to achieve high accuracy with standard loss functions. Recent studies have explored alternative loss functions, such as Contrastive Loss [18] and Focal Loss, to address these challenges. Contrastive Loss aims to enhance the discriminative power of the model by increasing the distance between different emotion classes in the feature space. Focal Loss, on the other hand, addresses class imbalance by focusing more on hard-to-classify examples.

This paper proposes a novel SER framework that combines the strengths of CNNs and Transformers while exploring the impact of different loss functions on recognition performance. By integrating CNNs for local feature extraction and Transformers for global context modeling, our approach aims to capture the intricate dynamics of emotional speech more effectively. Furthermore, we investigate the effectiveness of various loss functions, including Cross-Entropy Loss, Contrastive Loss, and Focal Loss, in improving the robustness and accuracy of emotion recognition.

The remainder of this paper is organized as follows: Section 2 reviews related work in SER, focusing on deep learning approaches and loss functions. Section 3 details the proposed CNN-Transformer architecture and the different loss functions explored. Section 4 discusses the results and compares the performance of different loss functions. Finally, Section 5 concludes the paper and outlines future research directions.

2. Related Work

The primary goal of SER is to extract emotion-related features from speech signals and classify them into predefined categories such as happiness, sadness, anger, fear, and neutrality. Due to the complexity and variability of speech signals, SER faces challenges like speaker variability, language differences, and environmental noise. The typical SER pipeline consists of the following steps:

Speech Signal Preprocessing: This includes noise reduction, framing, and windowing of the raw speech signal.

Feature Extraction: Extracting emotion-related features from the preprocessed speech signal. Traditional methods use handcrafted features, while deep learning methods automatically learn features from raw data.

Emotion Classification: Using machine learning or deep learning models to classify the extracted features into emotion categories.

Deep learning has revolutionized SER by enabling automatic feature extraction and improving classification accuracy. Two prominent deep learning architectures used in SER are Convolutional Neural Networks (CNNs) and Transformers.

The Transformer is an advanced deep learning architecture, first proposed for natural language processing (NLP). It relies entirely on attention mechanisms, specifically self-attention, eliminating the need for recurrent or convolutional structures. This design enables it to effectively model long-term dependencies and capture global context within sequential data. Transformers efficiently model long-range relationships in sequences through their multi-head self-attention mechanism. Unlike recurrent neural networks (e.g., RNN, LSTM), Transformers can directly and simultaneously consider interactions across all parts of the sequence.

The formula for Scaled Dot-Product Attention:

$$\text{Attention}\left(Q,K,V\right)=\text{softmax}\left(\frac{QK^{T}}{\sqrt{d_{k}}}\right)V$$

where, Q,K,V are Query, Key, and Value matrices, respectively.

3. Proposed Approach

Before giving the details of architecture, we illustrated our motivations.

3.1. Motivation

CNNs have characteristics that make them ideally suited for handling Mel-spectrograms in speech emotion recognition tasks: CNNs excel at extracting local features from two-dimensional inputs (such as images and spectrograms). Mel-spectrograms naturally resemble 2D images (time vs. frequency), allowing CNNs to leverage their strengths in spatial feature extraction. Local variations in pitch, tone, and intensity—critical indicators of emotion—are effectively captured by the convolutional filters of CNNs.

After CNNs extract essential local patterns, the Transformer module further enhances emotional representation by modeling long-range dependencies. Transformers use a mechanism called multi-head self-attention, originally popularized in natural language processing, to identify critical global relationships within the audio sequences.

Uses multiple self-attention layers simultaneously, each capturing different aspects of the emotional speech characteristics. Each head independently focuses on different subsets of the audio representation, allowing parallel extraction of diversified emotional information.

Integrating CNN and Transformer architectures creates a powerful hybrid system that leverages complementary strengths:

- CNN captures detailed local acoustic patterns essential for recognizing subtle emotional cues.
- Transformer effectively captures long-term dependencies and contextual interactions across the entire audio clip.

This integration significantly enhances the model's capability to understand both subtle short-term features (such as momentary pitch changes) and broader, long-term context, such as emotional tone evolution over speech duration.

3.2. Architecture

In this section, we illustrate the detailed pipeline for Speech Emotion Recognition using CNN-Transformer architecture, as shown in Figure 1. The process includes feature extraction, CNN processing, Transformer-based attention mechanisms, and ultimately the classification of emotions. Below is a detailed step-by-step explanation of the diagram.

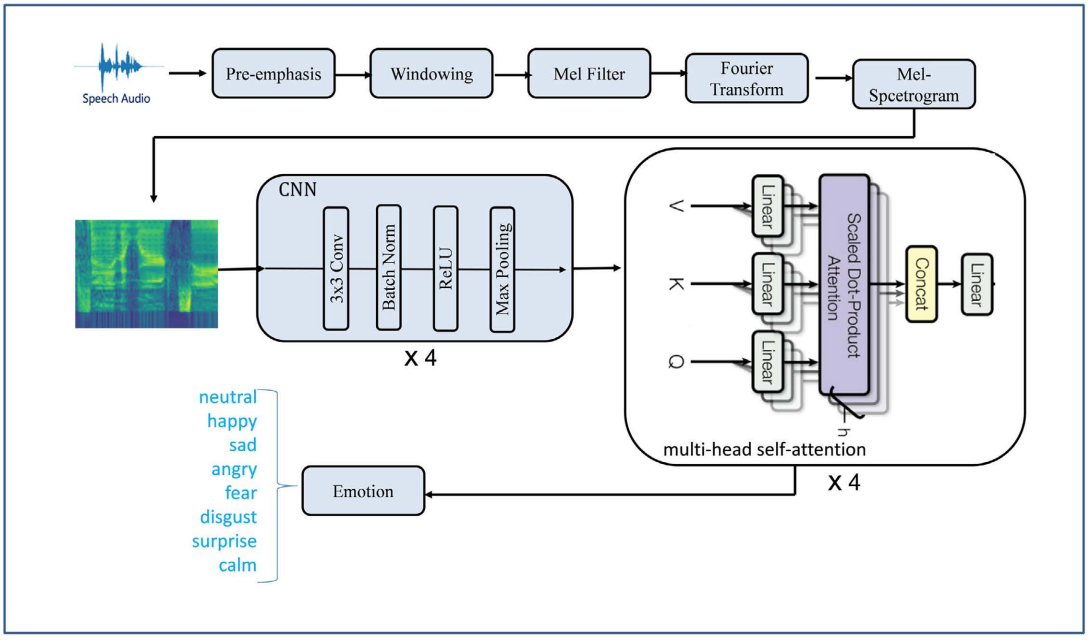


Figure 1. Pipeline of the proposed SER model.

First, the signal is input and extracted feature. The initial input is raw audio data (speech signal). The pre-emphasized signal is segmented into short frames using window functions (e.g., Hamming or Hanning window). Windowing facilitates the subsequent frequency analysis by reducing signal discontinuities. Each frame undergoes a Fourier Transform, converting the signal from the time domain to the frequency domain. Mel filters are applied to the frequency domain data, capturing perceptually relevant frequency bands, aligning with human auditory perception. The output of the Mel filtering process produces Mel-Spectrograms. These are visual representations capturing time-frequency information, crucial for detecting emotional nuances in speech signals.

Secondly, the Convolutional Neural Network (CNN) is used for further feature extraction. Especially, the Mel-spectrograms are input into a convolutional neural network, which applies convolutional filters to extract spatial features and important patterns. The CNN is employed multiple times (×4 layers or blocks) to progressively learn and refine discriminative emotional features from the Mel-Spectrogram. Specifically, as shown in **Figure 1**, each CNN block contains convolutional layers with kernel size 3, stride 1, and padding 1, followed by batch normalization and activation functions (ReLU). Max pooling layers progressively reduce spatial dimensions (kernel sizes of 2, 2, 2, and 4, respectively). The number of channels increases progressively from 16 to 64, capturing hierarchical features, while pooling layers progressively down sample the spatial dimensions, extracting high-level, compact representations from input data.

Thirdly, after CNN feature extraction, the data passes to a Transformer module, which leverages a multi-head self-attention mechanism (also repeated 4 times). The self-attention mechanism calculates attention scores based on query (Q), key (K), and value (V) matrices derived from the CNN outputs. Scaled dot-product attention is computed to weight the importance of different time-frequency regions effectively. Multiple attention heads (denoted as "multi-head") operate simultaneously to capture various aspects of the data from multiple representation subspaces. Outputs of different attention heads are concatenated and passed through a linear layer for dimensionality reduction and information fusion. In details, our Transformer includes an initial Max Pooling operation with kernel size [2, 4] and stride [2, 4] for dimensionality reduction before processing. The Transformer itself employs a model dimension typically consistent with the CNN's output channel size (64), includes 4 attention heads, and features internal feed-forward layers regulated by GELU activation. This design incorporates a low-dimensional, efficient representation, aided by the hyper parameters such as the intermediate feed-forward layer size (512) and multi-head self-attention structure, enabling the model to capture rich contextual information and effectively encode spatial-temporal patterns within the compressed input features.

The final Transformer output is fed into a classification layer that predicts the emotion category from the following classes: neutral, happy, sad, angry, fear, disgust, surprise, and calm.

This entire pipeline leverages both CNN's powerful feature extraction and Transformer's robust modeling of temporal dependencies, significantly enhancing the accuracy and reliability of emotion recognition from speech.

4. Experiments

4.1. Data and Setting

To strengthen the generalization performance of our method across varied datasets, we constructed a combined speech-emotion dataset using three prominent benchmarks: RAVDESS [19], SAVEE [20], and TESS [21]. Notably, these datasets differ in audio length and initial frames. Specifically, the RAVDESS dataset includes 1440 audio samples from 24 participants (equally distributed among females and males). SAVEE consists of 480 audio clips recorded by 4 male speakers, whereas TESS provides 2800 audio clips from 2 female speakers. Collectively, our unified dataset contains 4720 audio clips representing eight emotional states: calmness, happiness, sadness, anger, fear, surprise, disgust, and neutrality. While combining datasets with different distributions could potentially introduce biases or harmonization challenges, it also significantly enhances the diversity and representativeness of the resulting dataset.

Note that while combining datasets with different distributions could potentially introduce biases or harmonization challenges, it also significantly enhances the diversity and representativeness of the training samples. By integrating datasets collected under varying conditions (e.g., speaker identities, recording environments, and expressive styles), the unified dataset becomes more reflective of real-world variability. Furthermore, consistent preprocessing methods (such as silence removal using the librosa library and standardized audio segmentation) effectively mitigate harmonization challenges. Additionally, employing data augmentation techniques, like tripling each audio sample, further reduces any residual biases, strengthening the model's generalization capability. Overall, the merits of increased diversity and improved generalization through dataset combination substantially outweigh potential distributional biases.

For preprocessing, we utilized the "librosa" library to remove silence from each audio sample. A standardized 3.5-second segment was then extracted from the newly determined starting frame to ensure consistent duration across samples. We split the composite dataset into training, validation, and test subsets following an 80:10:10 distribution. To counteract potential overfitting, we augmented the training data by adding Gaussian noise to the original samples, thus tripling each clip.

When training the model, 4 different loss functions are employed, which are L1 loss, Smooth L1 loss, Binary Cross-Entropy loss, and Cross-Entropy loss. The L1 Loss measures the absolute difference between each predicted value and the actual target value. Essentially, it reflects the total magnitude of errors without considering direction. Due to its linear characteristic, L1 Loss tends to be robust against outliers—large errors do not disproportionately influence the outcome. It's fre-

quently used in regression tasks, particularly when robustness to outliers is desirable. Smooth L1 Loss combines advantages of L1 and L2 losses by behaving like L2 (mean squared error) when errors are small and behaving like L1 (mean absolute error) when errors are large. This adaptive behavior reduces sensitivity to outliers compared to the pure L2 loss, while still being differentiable everywhere, making optimization more stable.

Binary Cross-Entropy Loss measures the performance of a classification model whose output is a probability value between 0 and 1. It evaluates how well the predicted probability aligns with the true binary labels (0 or 1). If the predicted probability diverges significantly from the actual class, the penalty (loss) increases dramatically, incentivizing accurate predictions. Cross-Entropy Loss generalizes Binary Cross-Entropy Loss for multiple-class classification scenarios. It evaluates how effectively predicted class probabilities match the true class labels, which are typically represented as one-hot vectors (only one correct class per input). A good prediction (probability close to 1 for the correct class and close to 0 for incorrect ones) yields a low loss; a poor prediction yields a high loss.

Our choice of these losses was motivated by their suitability for different tasks: L1 and Smooth L1 losses were selected for regression-like tasks requiring robustness to outliers, while Binary Cross-Entropy Loss naturally fits binary classification scenarios involving probability predictions.

4.2. Results and Analysis

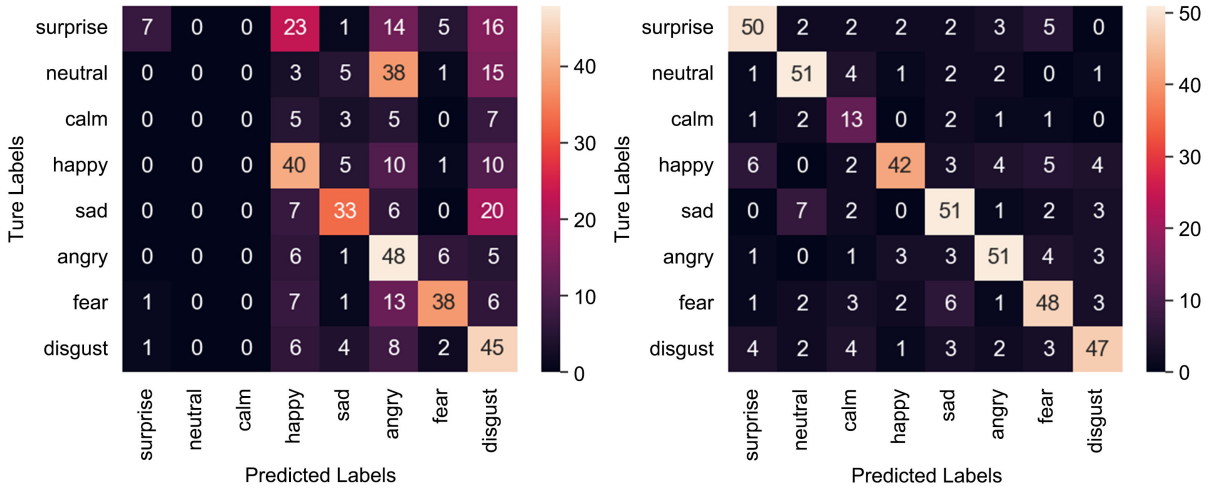
The comparison of the confusion matrices is shown in Figure 2. The confusion matrices clearly show the impact of different loss functions on the performance of the speech emotion recognition model.

The L1 Loss yields the poorest results, exhibiting significant confusion among classes, particularly misclassifying many emotions as "happy" or "disgust." Smooth L1 Loss provides improved accuracy compared to L1 Loss, significantly reducing misclassifications, though some confusion remains among closely related emotional states. Binary Cross-Entropy Loss further enhances classification accuracy, clearly distinguishing most emotions but still showing minor confusion between similar classes. Cross-Entropy Loss demonstrates the best overall performance, achieving the highest accuracy with minimal misclassifications, effectively distinguishing between emotional categories with minimal overlap. Overall, the Cross-Entropy Loss offers superior performance in accurately classifying speech emotions, indicating its suitability for multiclass emotion recognition tasks.

Figure 3 shows the loss comparison of the model in different loss functions. Overall, Cross-Entropy Loss provides the most stable, fastest convergence, and best generalization performance, while L1 Loss performs poorly. Smooth L1 and Binary Cross-Entropy show intermediate performance with some instability or mild over-fitting compared to Cross-Entropy Loss.

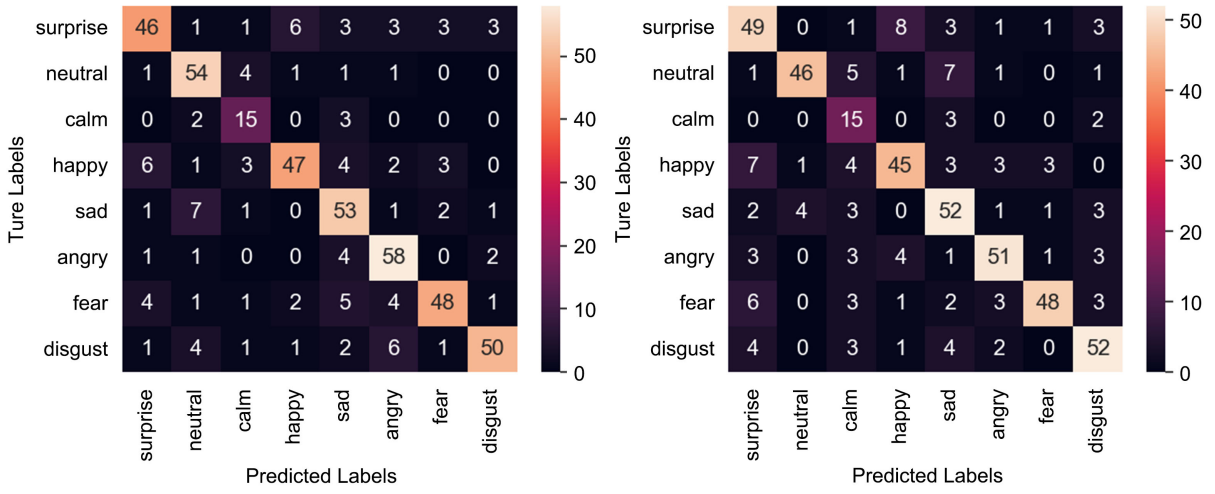
Smooth L1 Loss demonstrates stable and gradual convergence, with both training and validation losses decreasing steadily over epochs. The validation loss sta-

bilizes over time, indicating consistent and reliable learning. Binary Cross-Entropy Loss presents rapid initial convergence in training loss but reveals a noticeable gap between training and validation losses. Although generally stable, some minor oscillations in validation loss indicate potential slight over fitting. Cross-Entropy Loss exhibits fast convergence with highly stable and low final values for both training and validation losses. The minimal gap between the two suggests the best generalization and most effective learning among all methods.



(a) L1 Loss

(b) Smooth L1 Loss



(c) Binary Cross-Entropy Loss

(d) Cross-Entropy Loss

Figure 2. Confusion matrices of the model in different loss functions.

Figure 4 shows that the recognition accuracy compares of the model in different loss functions. Among the four loss functions, the Cross-Entropy Loss clearly delivers the best balance between model performance, convergence speed, and generalization capability. Binary Cross-Entropy Loss performs relatively well but shows slight over fitting. Smooth L1 Loss is better than L1 but still lags behind Cross-Entropy Loss. L1 Loss performs poorly in terms of stability and overall ac-

curacy. Therefore, for speech emotion recognition tasks, Cross-Entropy Loss appears to be the most effective and recommended choice, providing the best balance between convergence speed, accuracy, and model generalization.

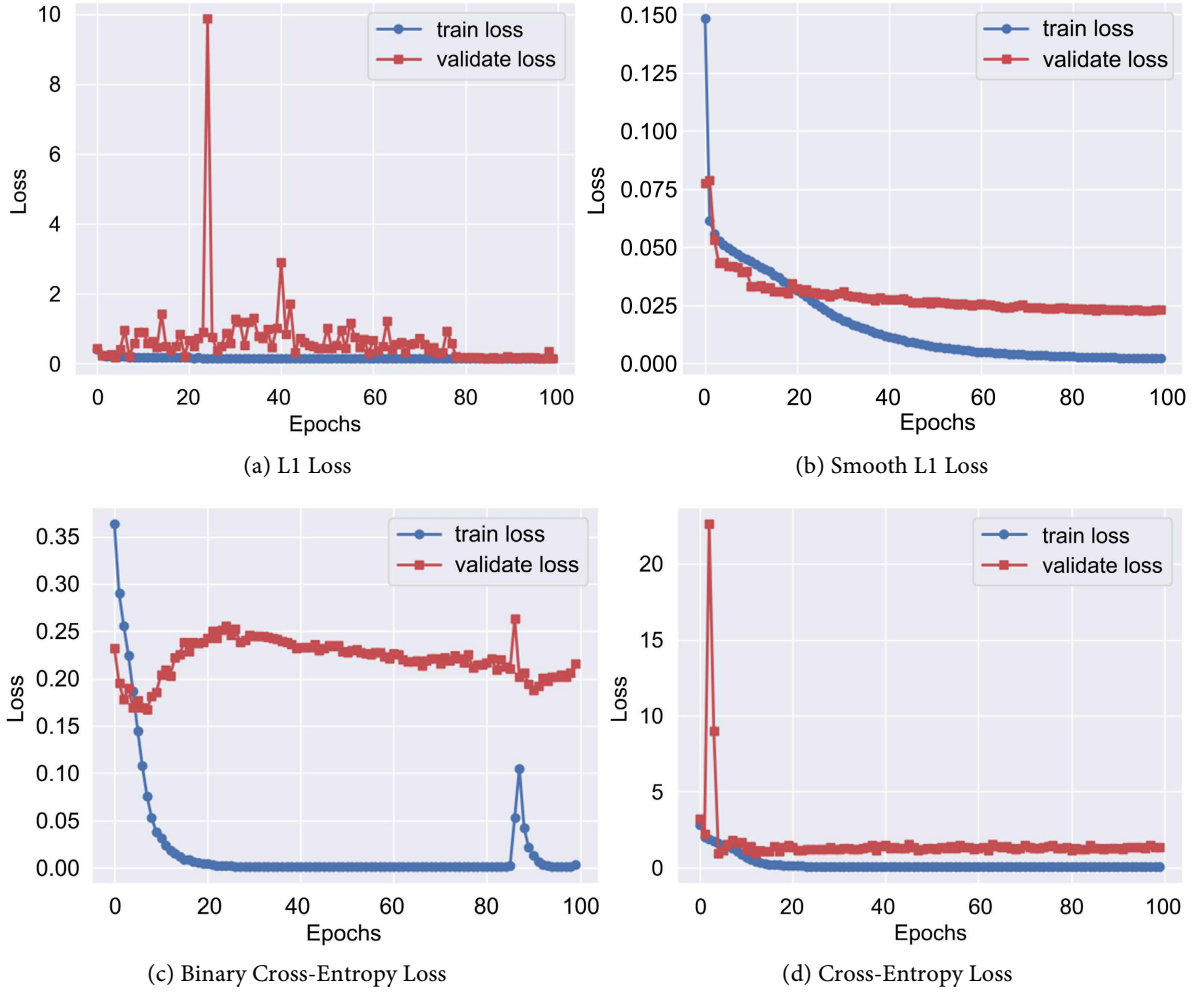
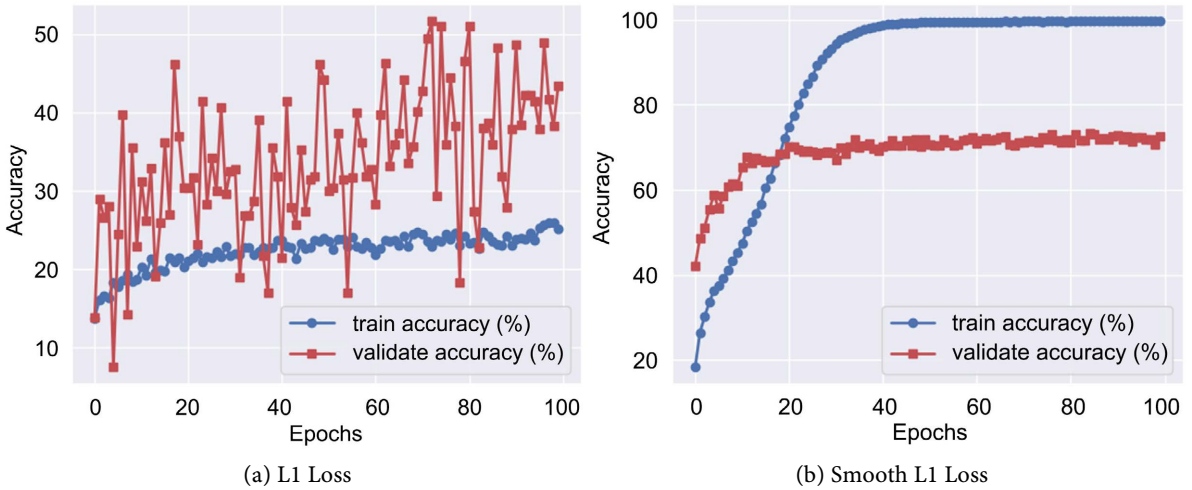
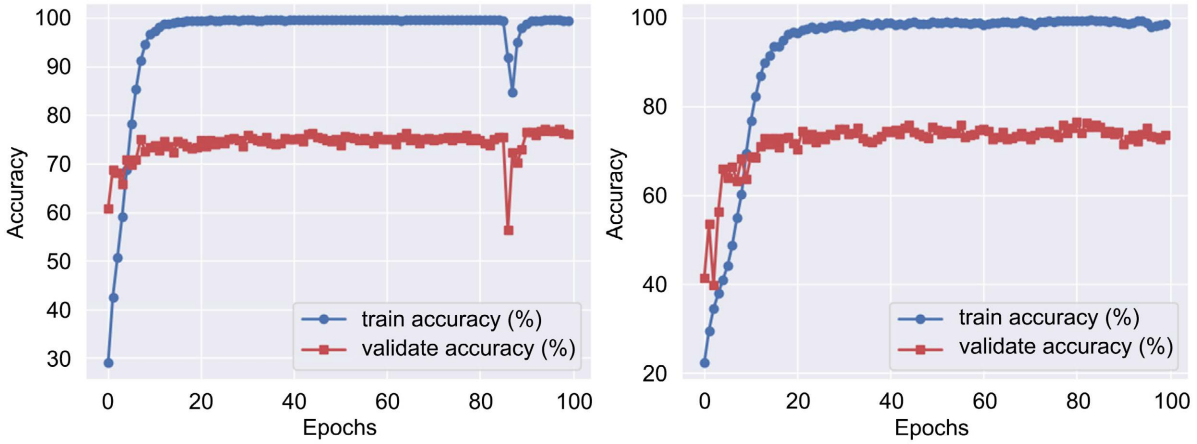


Figure 3. Loss compare of the model in different loss functions.





(c) Binary Cross-Entropy Loss (d) Cross-Entropy Loss

Figure 4. Recognition accuracy compares of the model in different loss functions.

Table 1 presents the overall performance of the model trained with different loss functions. L1 Loss demonstrates poor convergence and accuracy, making it the least suitable option for speech emotion recognition. Smooth L1 Loss significantly improves accuracy but shows mild signs of over fitting. Binary Cross-Entropy Loss exhibits good generalization; however, the validation loss still suggests potential over fitting. In contrast, Cross-Entropy Loss stands out as the most effective, achieving the highest validation and test accuracy. Therefore, it is recommended for robust and reliable speech emotion recognition tasks.

Table 1. The overall performance of the model in different loss functions.

Loss function	Loss Values			Recognition Accuracy (%)		
	Train	Validate	Test	Train	Validate	Test
L1	0.120	0.142	0.124	28.12	43.31	44.14
Smooth L1	0.003	0.023	0.021	100	72.61	73.85
Binary Cross-Entropy	0.002	0.215	0.198	100	76.01	77.62
Cross-Entropy	0.003	1.283	1.153	100	73.46	74.90

Table 2 summarizes the Precision, Recall, and F1-score performance on the test set. While L1 Loss yields relatively high precision (0.6357), it suffers from notably low recall (0.3954) and F1-score (0.4233). The other three loss functions—Smooth L1, Binary Cross-Entropy, and Cross-Entropy—deliver more balanced results. Among them, Binary Cross-Entropy achieves the best overall performance (Precision: 0.6275, Recall: 0.6255, F1-score: 0.6238). Cross-Entropy Loss follows closely, demonstrating consistent and stable behavior across all metrics, reinforcing its effectiveness in multiclass classification scenarios. Overall, Binary Cross-Entropy offers the best combination of precision, recall, and F1-score among the evaluated loss functions.

Table 2. Comparison of Precision, Recall, and F1-score across different loss functions.

Loss function	Precision	Recall	F1-score
L1	0.6357	0.3954	0.4233
Smooth L1	0.6102	0.6109	0.6069
Binary Cross-Entropy	0.6275	0.6255	0.6238
Cross-Entropy	0.6198	0.6130	0.6077

5. Conclusion

This paper proposed a CNN-Transformer hybrid architecture for speech emotion recognition (SER), effectively integrating CNN's capability in extracting local features and Transformer's strength in modeling global contextual dependencies. Extensive comparative experiments on various loss functions demonstrated that Cross-Entropy Loss significantly outperformed other alternatives, achieving superior convergence, higher accuracy, and enhanced generalization across validation and test scenarios. The superior performance of Cross-Entropy Loss underscores its suitability for robust multiclass SER tasks. From a practical viewpoint, the demonstrated effectiveness of this hybrid approach facilitates deployment in real-world scenarios requiring reliable emotional understanding, such as virtual assistants, human-computer interaction, healthcare monitoring systems, and emotional state tracking in automated services. Future work should further explore lightweight variants of the architecture and investigate adaptability across diverse datasets and realistic deployment conditions, ensuring broader applicability and scalability in technology applications.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Al-Dujaili, M.J. and Ebrahimi-Moghadam, A. (2023) Speech Emotion Recognition: A Comprehensive Survey. *Wireless Personal Communications*, **129**, 2525-2561. https://doi.org/10.1007/s11277-023-10244-3

[2] Singh, Y.B. and Goel, S. (2022) A Systematic Literature Review of Speech Emotion Recognition Approaches. *Neurocomputing*, **492**, 245-263. https://doi.org/10.1016/j.neucom.2022.04.028

[3] Ancilin, J. and Milton, A. (2021) Improved Speech Emotion Recognition with Mel Frequency Magnitude Coefficient. *Applied Acoustics*, **179**, Article 108046. https://doi.org/10.1016/j.apacoust.2021.108046

[4] Warule, P., Mishra, S.P., Deb, S. and Krajewski, J. (2023) Sinusoidal Model-Based Diagnosis of the Common Cold from the Speech Signal. *Biomedical Signal Processing and Control*, **83**, Article 104653. https://doi.org/10.1016/j.bspc.2023.104653

[5] Zhao, X., Zhang, S. and Lei, B. (2013) Robust Emotion Recognition in Noisy Speech via Sparse Representation. *Neural Computing and Applications*, **24**, 1539-1553. https://doi.org/10.1007/s00521-013-1377-z

[6] Jothimani, S. and Premalatha, K. (2022) MFF-SAug: Multi Feature Fusion with Spectrogram Augmentation of Speech Emotion Recognition Using Convolution Neural Network. *Chaos, Solitons & Fractals*, **162**, Article 112512. https://doi.org/10.1016/j.chaos.2022.112512

[7] Issa, D., Fatih Demirci, M. and Yazici, A. (2020) Speech Emotion Recognition with Deep Convolutional Neural Networks. *Biomedical Signal Processing and Control*, **59**, Article 101894. https://doi.org/10.1016/j.bspc.2020.101894

[8] Aftab, A., Morsali, A., Ghaemmaghami, S. and Champagne, B. (2022) LIGHT-SERNET: A Lightweight Fully Convolutional Neural Network for Speech Emotion Recognition. *ICASSP* 2022-2022 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Singapore, 23-27 May 2022, 6912-6916. https://doi.org/10.1109/icassp43922.2022.9746679

[9] Alluhaidan, A.S., Saidani, O., Jahangir, R., Nauman, M.A. and Neffati, O.S. (2023) Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network. *Applied Sciences*, **13**, Article 4750. https://doi.org/10.3390/app13084750

[10] Liu, Z., Han, M., Wu, B. and Rehman, A. (2023) Speech Emotion Recognition Based on Convolutional Neural Network with Attention-Based Bidirectional Long Short-Term Memory Network and Multi-Task Learning. *Applied Acoustics*, **202**, Article 109178. https://doi.org/10.1016/j.apacoust.2022.109178

[11] Hazmoune, S. and Bougamouza, F. (2024) Using Transformers for Multimodal Emotion Recognition: Taxonomies and State of the Art Review. *Engineering Applications of Artificial Intelligence*, **133**, Article 108339. https://doi.org/10.1016/j.engappai.2024.108339

[12] Zhang, S., Liu, R., Yang, Y., Zhao, X. and Yu, J. (2022) Unsupervised Domain Adaptation Integrating Transformer and Mutual Information for Cross-Corpus Speech Emotion Recognition. *Proceedings of the* 30*th ACM International Conference on Multimedia*, Lisboa, 10-14 October 2022, 120-129. https://doi.org/10.1145/3503161.3548328

[13] Swain, M., Maji, B., Khan, M., Saddik, A.E. and Gueaieb, W. (2023) Multilevel Feature Representation for Hybrid Transformers-Based Emotion Recognition. 2023 5*th International Conference on Bio-Engineering for Smart Technologies* (*BioSMART*), Paris, 7-9 June 2023, 1-5. https://doi.org/10.1109/biosmart58455.2023.10162089

[14] Hu, K., Xu, K., Xia, Q., Li, M., Song, Z., Song, L., *et al.* (2024) An Overview: Attention Mechanisms in Multi-Agent Reinforcement Learning. *Neurocomputing*, **598**, Article 128015. https://doi.org/10.1016/j.neucom.2024.128015

[15] Alzubaidi, L., Zhang, J., Humaidi, A.J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., *et al.* (2021) Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions. *Journal of Big Data*, **8**, Article No. 53. https://doi.org/10.1186/s40537-021-00444-8

[16] Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., *et al.* (2021) A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges. *Information Fusion*, **76**, 243-297. https://doi.org/10.1016/j.inffus.2021.05.008

[17] Mao, A., Mohri, M. and Zhong, Y. (2023) Cross-Entropy Loss Functions: Theoretical Analysis and Applications. *Proceedings of the* 40*th International Conference on Machine Learning*, Honolulu, 23-29 July 2023, 23803-23828.

[18] Wang, F. and Liu, H. (2021) Understanding the Behaviour of Contrastive Loss. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 2495-2504. https://doi.org/10.1109/cvpr46437.2021.00252

[19] Livingstone, S.R. and Russo, F.A. (2018) 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*, **13**, e0196391. https://doi.org/10.1371/journal.pone.0196391

[20] Jackson, P. and Haq, S. (2015) Surrey Audio-Visual Expressed Emotion (SAVEE) Database. http://kahlan.eps.surrey.ac.uk/savee/

[21] Dupuis, K. and Pichora-Fuller, M.K. (2010) Toronto Emotional Speech Set (TESS). https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess