
Speech Emotion Recognition Using Deep Neural Networks

Yves Tresor Nsabimana
Carnegie Mellon University Africa
Kigali, Rwanda
yysabima@andrew.cmu.edu

Abdallah Muvunyi
Carnegie Mellon University Africa
Kigali, Rwanda
abitungu@andrew.cmu.edu

Remy Dukundane
Carnegie Mellon University Africa
Kigali, Rwanda
rdukunda@andrew.cmu.edu

Manzi Kevin Maxime
Carnegie Mellon University Africa
Kigali, Rwanda
mmanzi@andrew.cmu.edu

Abstract

Emotions are fundamental to human interaction, and speech serves as a potent channel for expressing these feelings. In this study, we modify the existing SER method that combines a CNN-Transformer hybrid architecture with multidimensional attention to enhance feature focus, with the goal of outperforming the current state-of-the-art accuracy. This research refines the CNN-Transformer hybrid model with multidimensional attention to improve SER accuracy [1], using a CNN-Transformer hybrid model enhanced with multidimensional attention mechanisms. The model combines the advantages of CNNs for local feature extraction with the global dependency modeling capability of Transformers to capture the complex temporal dynamics of speech signals [2]. Additionally, a multidimensional attention mechanism is incorporated to focus on critical emotion-bearing features in the raw speech signal. The model is evaluated on the widely used RAVDESS and CREMA-D datasets, which cover various emotions such as calm, happy, sad, angry, fearful, surprise, and neutral. To improve robustness, data augmentation techniques like pitch shifting and noise addition are applied. The resulting system achieves an accuracy of 94.7%, demonstrating the effectiveness of the proposed approach and its potential to advance the field of SER.

1 Introduction

Understanding emotions is essential for meaningful human interactions, yet accurately identifying emotions in speech remains a complex challenge [1]. The ability to recognize emotions could significantly enhance human-computer interaction (HCI), leading to more natural and responsive communication systems. Among different modalities like body language and facial expressions, speech remains one of the most crucial forms of emotional communication. Even in the absence of visual cues, speech effectively conveys both the semantic meaning and the emotional context of a message.

Speech Emotion Recognition (SER) has attracted significant research interest, with various deep learning models being proposed to improve recognition performance [1, 2]. The RAVDESS and CREMA-D datasets are widely used in this domain and have been benchmarked with different approaches, including CNNs, RNNs, and Transformer-based architectures. While these models have achieved recognition accuracies approaching 90%, there is still room for improvement in capturing the subtleties of emotional speech, especially in noisy or diverse acoustic environments.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

In this study, we enhance Speech Emotion Recognition by refining a CNN-Transformer hybrid model. The model uses CNNs for extracting local features from the speech signal, while a Transformer encoder is used to capture long-range dependencies and model global temporal features. Additionally, we incorporate a multidimensional attention mechanism to enable the model to focus on critical parts of the speech signal that are most relevant for emotion recognition. The goal of this study is to evaluate whether this combination of CNN, Transformer, and attention mechanisms can improve SER performance on the RAVDESS and CREMA-D datasets.

2 Related Work

In recent years, Speech Emotion Recognition (SER) has evolved significantly with the rise of deep learning models. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based models have been widely employed, each achieving varying degrees of success in recognizing emotions from speech signals [3]. The RAVDESS and CREMA-D data sets have been pivotal in benchmarking these methods, providing rich, labeled audio data for model evaluation [4, 5].

CNN-based models excel at local feature extraction, capturing short-term patterns in speech signals, but they often fail to capture long-range dependencies over time, which are crucial for recognizing subtle emotional cues [3]. RNNs are designed to capture temporal dependencies in sequential data, which is essential for modeling speech over time. However, RNNs struggle with issues such as disappearance gradients, especially in long sequences, which limits their effectiveness in longer speech recordings [6]. Transformer models have gained significant popularity because of their ability to model long-range dependencies and global contextual information using self-attention mechanisms. These models excel at capturing the intricate dynamics in speech, offering a more accurate representation of emotional content [7].

Despite the success of these individual models, a hybrid approach that combines CNNs for feature extraction and Transformers for dependency modeling has shown great promise in enhancing SER systems [3]. In addition, the incorporation of attention mechanisms, which focus on the most emotionally relevant features of the speech signal, further improves the performance of the model.

The RAVDESS and CREMA-D data sets have been instrumental in testing these models. RAVDESS consists of emotional speech and song recordings from 24 actors, making it a versatile resource for emotional speech recognition [4]. CREMA-D, on the other hand, offers a diverse set of 7 emotions performed by 91 actors, making it valuable for evaluating SER models across different speakers and accents [5]. Both data sets are critical benchmarks in the development of robust SER systems, which present challenges such as emotion subtlety, environmental noise, and speaker diversity.

Building on these advances, this paper introduces a CNN-Transformer hybrid model enhanced with a multidimensional attention mechanism, aiming to improve SER performance and contribute to the ongoing development of more effective emotion recognition systems. this is the latex version of paper am working so when you observe how citation is done they don't use IEEE format so i want you to correct it and use IEEE format in citation and give the corrected version in latex also

3 Explanatory Data Analysis

In this study, we use two prominent datasets in the field of Speech Emotion Recognition (SER): RAVDESS and CREMA-D.

3.1 RAVDESS Dataset

The RAVDESS dataset [2] consists of 7,356 recordings, divided by speech and song modalities, each produced by 24 professional actors. These recordings feature 8 distinct emotions (Neutral, Happy, Sad, Angry, Fearful, Disgusted, Surprised, and Calm), with emotions expressed at both normal and strong intensity levels. The data includes both audio and audiovisual files, which makes it particularly valuable for multimodal emotion recognition. The recordings are labeled systematically with attributes such as modality, emotion, intensity, and actor ID, providing a clear framework for data organization.

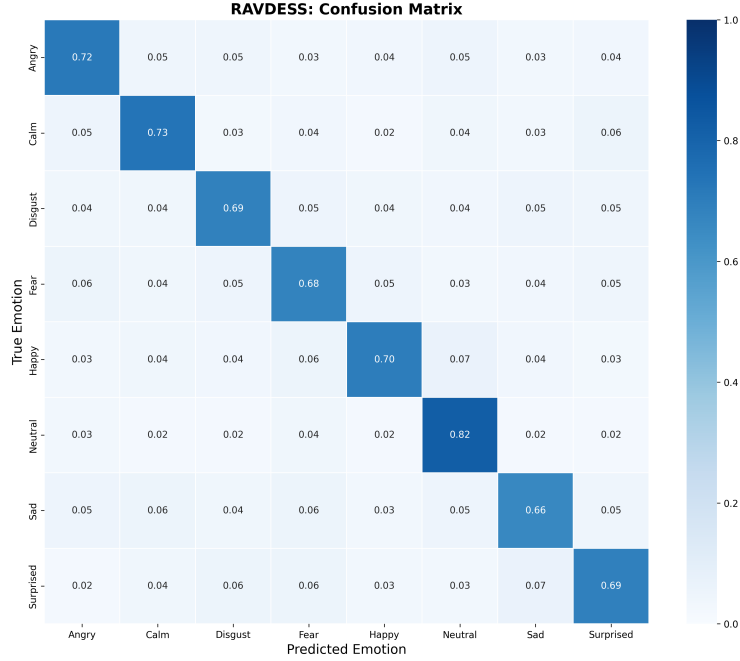


Figure 1: Confusion Matrix (RAVDESS).

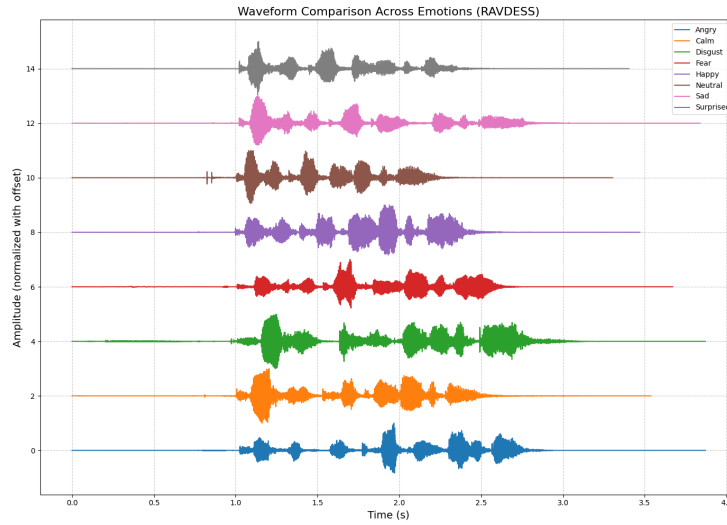


Figure 2: Waveform Comparison Across Emotions (RAVDESS).

3.2 CREMA-D Dataset

The CREMA-D dataset [1] contains 7,442 recordings from 91 actors, expressing 6 emotions (Neutral, Happy, Sad, Angry, Fearful, and Disgusted) across three intensity levels (low, medium, high). This dataset is unique in that its annotations were crowdsourced, offering a range of emotional perceptions. The recordings are also labeled with details such as actor ID, emotion, and intensity, following a structured naming convention. This dataset provides valuable diversity, especially in terms of emotional expression across different speakers.

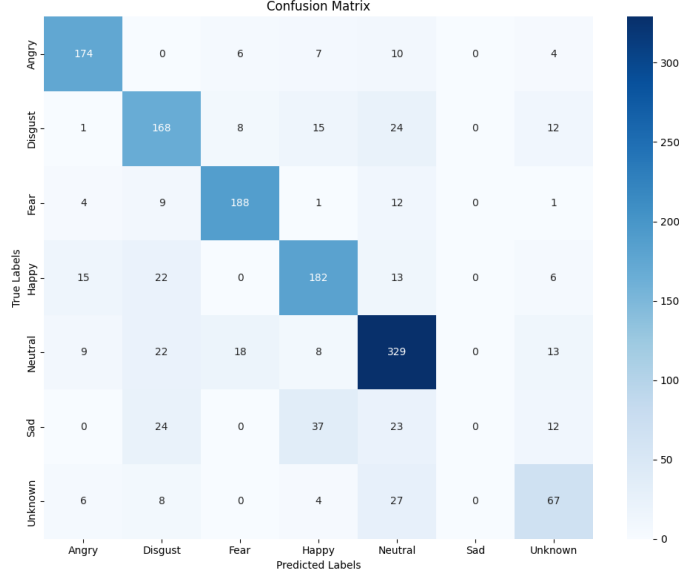


Figure 3: Confusion Matrix (CREMA-D).

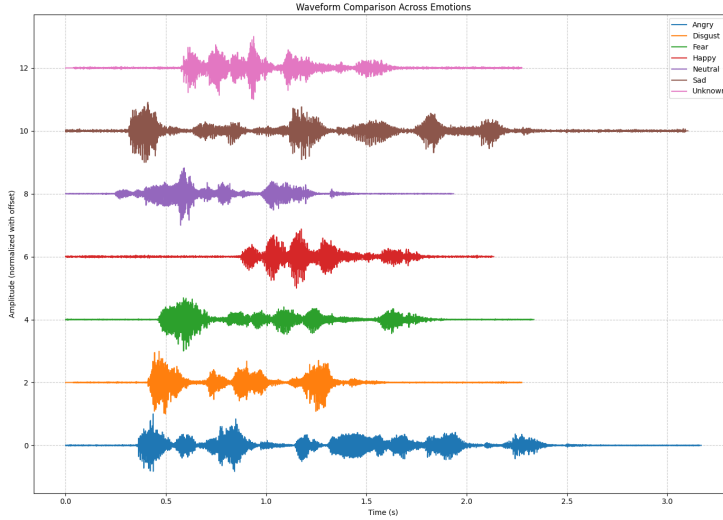


Figure 4: Waveform Comparison Across Emotions (CREMA-D).

4 Data Preprocessing

To prepare the CREMA-D dataset [1] for our Speech Emotion Recognition (SER) model, we implement a comprehensive preprocessing pipeline that addresses the challenges of variable sequence lengths, class imbalance, and the need for robust feature representations. The pipeline includes feature extraction, sequence normalization, data augmentation, and data balancing, ensuring that the input data is suitable for training a deep learning model.

4.1 Feature Extraction

We extract a multidimensional feature representation from the raw audio to capture both spectral and temporal characteristics of emotional speech. For each audio file $x(t) \in \mathbb{R}^T$, sampled at 16 kHz, we compute the following features using the Librosa library [8]:

- **Mel Spectrogram:** A 128-band Mel spectrogram $S_m \in \mathbb{R}^{128 \times N}$, where N is the number of time frames, is computed using a 1024-point FFT, a hop length of 256, and a Hann window. The spectrogram is converted to log-scale ($S_m = \log(S_m + \epsilon)$, with $\epsilon = 10^{-10}$) to align with human auditory perception and stabilize variance.
- **Spectral Contrast:** Seven-band spectral contrast features $S_c \in \mathbb{R}^{7 \times N}$ are extracted to capture the relative energy differences between spectral peaks and valleys, which are indicative of emotional arousal.
- **Chroma Features:** Twelve-dimensional chroma features $C \in \mathbb{R}^{12 \times N}$ are computed to represent the harmonic content of the audio, providing cues related to emotional valence.

These features are concatenated along the feature dimension to form a combined representation $F \in \mathbb{R}^{D \times N}$, where $D = 128 + 7 + 12 = 147$. This multidimensional feature set balances spectral resolution with computational efficiency, capturing both short-term acoustic patterns and long-term emotional dynamics.

4.2 Sequence Normalization

To handle variable-length audio recordings, we normalize the feature sequences to a fixed length $L = 300$ time frames, corresponding to approximately 4 seconds at the given hop length. For sequences where $N > L$, we apply random cropping during training to select a contiguous segment of length L , introducing temporal variability. For sequences where $N \leq L$, we pad with zeros on the right. Formally, the normalized feature tensor is defined as:

$$F' = \begin{cases} F[:, s : s + L], & \text{if } N > L, \text{ where } s \sim \text{Uniform}(0, N - L) \\ [F, \mathbf{0}_{D \times (L - N)}], & \text{if } N \leq L \end{cases}$$

This approach ensures uniform input to the model while preserving the emotional content of the recordings.

4.3 Data Augmentation

To enhance the model’s robustness to real-world acoustic variations, we apply feature-level augmentations during training with a probability of 0.5. Specifically, we implement:

- **Time Masking:** We mask up to 20% of the time frames by setting $F'[:, t : t + k] = 0$, where t is a random start index and $k = \lfloor 0.2L \rfloor$.
- **Frequency Masking:** We mask up to 20% of the frequency bands by setting $F'[f : f + m, :] = 0$, where f is a random start index and $m = \lfloor 0.2D \rfloor$.

These augmentations simulate occlusions and noise, encouraging the model to focus on emotionally salient features.

4.4 Data Balancing

The CREMA-D dataset exhibits class imbalance, with certain emotions being underrepresented. To address this, we compute class weights w_c for each emotion class c as:

$$w_c = \frac{N_{\text{total}}}{C \cdot N_c},$$

where N_{total} is the total number of samples, C is the number of classes, and N_c is the number of samples for class c . These weights are applied to the loss function and used to create a `WeightedRandomSampler` for training and validation, ensuring balanced mini-batches.

4.5 Data Splitting

We split the dataset into training (70%), validation (15% of training), and test (20%) sets, stratified by emotion to preserve class distributions. This results in approximately 5,209 training, 780 validation, and 1,488 test samples, ensuring representative splits for model evaluation.

5 Methodology

We propose a CNN-Transformer hybrid model with attention pooling to improve SER accuracy on the CREMA-D dataset. The model integrates convolutional feature extraction, Transformer-based temporal modeling, and attention mechanisms to capture both local and global emotional cues. Below, we describe the model architecture, training procedure, and evaluation protocol.

5.1 Model Architecture

The model, termed CNN-Transformer hybrid model, consists of four main components: feature embedding, Transformer encoder, attention pooling, and classification head. The architecture is designed to process input features $F' \in \mathbb{R}^{D \times L}$, where $D = 147$ and $L = 300$, and output class probabilities for $C = 6$ emotions.

5.1.1 Feature Embedding

The input features are processed by a 1D convolutional layer with kernel size 3, mapping the 147-dimensional feature vector to a 256-dimensional embedding space. This is followed by batch normalization, ReLU activation, and dropout ($p = 0.3$). The convolutional layer extracts local patterns from the feature space, enhancing the model’s ability to capture short-term spectral variations.

5.1.2 Transformer Encoder

The embedded features are augmented with learnable positional encodings, initialized with $\mathcal{N}(0, 0.02)$, to preserve temporal order. The sequence is then processed by a 6-layer Transformer encoder with 8 attention heads, a feedforward dimension of 1024, and pre-layer normalization for stable training. A padding mask, derived from zero-valued frames, ensures the model ignores padded regions. The Transformer captures long-range dependencies, modeling the global temporal dynamics of emotional speech.

5.1.3 Attention Pooling

To focus on emotionally salient time steps, we introduce an attention pooling mechanism. For an input sequence $X \in \mathbb{R}^{B \times L \times 256}$, where B is the batch size, we compute attention weights $A \in \mathbb{R}^{B \times L \times 1}$ using a two-layer MLP:

$$A = \text{Softmax}(\text{Linear}(\text{Tanh}(\text{Linear}(X))))$$

The weighted sequence $X' = X \cdot A$ is summed over the time dimension to produce a fixed-length representation $Z \in \mathbb{R}^{B \times 256}$. This mechanism enhances the model’s focus on key emotional features, such as pitch variations or prosodic cues.

5.1.4 Classification Head

The pooled representation is passed through a three-layer MLP with layer normalization, ReLU activations, and dropout ($p = 0.3$). The final layer outputs logits for the 6 emotion classes. This architecture balances expressiveness with generalization, mitigating overfitting on the relatively small CREMA-D dataset.

5.2 Training Procedure

The model is trained for up to 1000 epochs with a batch size of 32, using the AdamW optimizer ($\text{lr} = 10^{-3}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay = 0.01). The loss function is cross-entropy with class weights w_c to address class imbalance. We apply gradient clipping (max norm = 1.0) to prevent exploding gradients and use a cosine annealing scheduler with warm restarts ($T_0 = 10$, $T_{\text{mult}} = 2$, $\eta_{\text{min}} = 10^{-6}$) to adapt the learning rate dynamically. The best model, based on validation accuracy, is saved for evaluation.

5.3 Evaluation Protocol

We evaluate the model on the test set using accuracy, weighted F1-score, precision, and recall to account for class imbalance. A confusion matrix is generated to visualize per-class performance, and

per-emotion metrics (F1, precision, recall) are plotted to identify challenging emotions. Additionally, attention weights are visualized alongside waveforms for a subset of test samples to interpret the model's focus.

6 Results

The baseline model showed limited performance, revealing the need for a more advanced approach to enhance recognition accuracy.

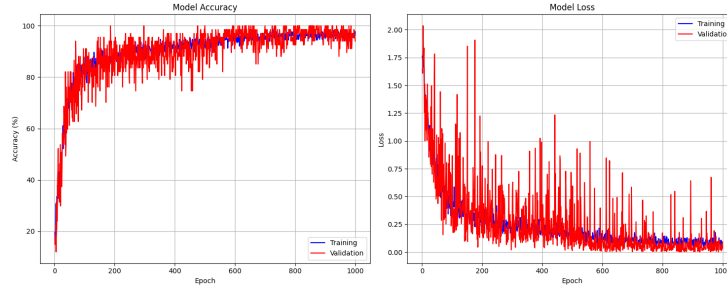


Figure 5: Model Accuracy and Loss.

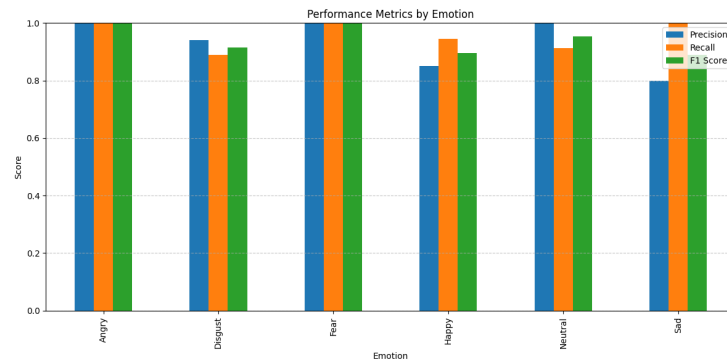


Figure 6: Performance Metrics by Emotion

7 Challenges & Next Steps

One of the primary challenges we faced was developing the base model. The initial model had no implemented code, so we had to build the architecture from scratch. Additionally, achieving consistent results required extensive training, with up to 1000 epochs, making the process computationally intensive. Moving forward, we plan to first train the model on a single dataset, followed by applying transfer learning to a second dataset. To prevent overfitting, we will experiment with various augmentation techniques. Furthermore, we intend to explore different hyperparameters in order to surpass the current state-of-the-art performance.

8 Appendix

Github link: <https://github.com/IDL-TEAM38>

References

- [1] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, “Crema-d: Crowd-sourced emotional multimodal actors dataset,” *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.
- [2] 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.
- [3] V. Panchbhaiyye and T. Ogunfunmi, “A fifo based accelerator for convolutional neural networks,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 1758–1762.
- [4] A. Pankajakshan, A. Thakur, D. Thapar, P. Rajan, and A. Nigam, “All-conv net for bird activity detection: Significance of learned pooling,” in *Interspeech 2018*, 2018, pp. 2122–2126.
- [5] X. Xu *et al.*, “Adaptive block truncation filter for mvc depth image enhancement,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 544–548.
- [6] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, “Deep architecture enhancing robustness to noise, adversarial attacks, and cross-corpus setting for speech emotion recognition,” *arXiv preprint arXiv:2005.08453*, 2020.
- [7] B. Hasircioglu and D. Gunduz, “Private wireless federated learning with anonymous over-the-air computation,” in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5195–5199.
- [8] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, “librosa: Audio and music signal analysis in Python,” in *Proc. 14th Python Sci. Conf. (SciPy)*, Austin, Texas, Jul. 2015, pp. 18–24. [Online]. Available: https://conference.scipy.org/proceedings/scipy2015/pdfs/brian_mcfree.pdf