Speech Emotion Recognition Analysis Using Deep Learning

Tasmiah Binte Azad
ID: 21225103457

Md Ashikur Rahman
ID: 21225103471

Humayra Kabir Hride
ID: 21225103501

Abstract

Speech is a powerful medium through which humans communicate not only information but also emotions that influence interaction, decision making, and social connection. With the increasing presence of intelligent systems in education, healthcare, and everyday technology, the ability for machines to accurately recognize emotions from speech is becoming a critical research area. Yet, current speech emotion recognition approaches often struggle to achieve reliable accuracy across multiple emotion categories, particularly when trained on limited datasets. Many models are effective at extracting short term acoustic cues but fail to capture the longer contextual patterns that shape how emotions are expressed in natural speech. This project will address that gap by developing and evaluating two hybrid deep learning models: one combining convolutional neural networks with long short term memory layers, and another combining convolutional neural networks with transformer based layers. Using the Toronto Emotional Speech Set, which provides recordings labeled across seven distinct emotions, the system will preprocess the audio data, extract features such as mel spectrograms, train the two models, and assess their performance using accuracy and F1 score. The expected outcome is a clearer understanding of which hybrid approach better models emotional information in speech, offering valuable insights for future emotion aware systems in human computer interaction, adaptive learning platforms, and mental health support tools.

1 Introduction and Motivation

Emotions are central to human communication, shaping interactions and decision-making. With the increasing adoption of intelligent systems, enabling machines to recognize emotions has become a vital research area. Speech Emotion Recognition (SER) allows computers to analyze vocal cues and classify emotions such as happiness, anger, sadness, or fear.

The motivation for this work stems from three major observations:

- **Growing importance of emotion-aware systems:** From mental health monitoring to conversational agents, emotion recognition has the potential to improve personalization, empathy, and decision-making.
- **Limitations of existing models:** Many current SER models achieve high accuracy on controlled datasets but fail in real-world noisy environments due to limited dataset diversity and scalability challenges.

- **Advances in deep learning:** Hybrid architectures (CNN-LSTM, CNN-Transformer, CLDNN) combined with feature extraction techniques (MFCCs, spectrograms) and augmentation strategies have demonstrated promising results in recent research.

This project is motivated by the need to build a generalizable and scalable SER framework capable of performing well across multiple languages, datasets, and conditions.

2 Literature Review

Research in speech emotion recognition (SER) has evolved through diverse datasets, architectures, and evaluation strategies, yet challenges remain in scalability, generalization, and real-world applicability. Shaik Abdul Khalandar Basha [1] employed CNN-LSTM with attention on RAVDESS, CREMA-D, and TESS, attaining 87.08% accuracy but facing dataset size and real-time deployment limitations. Similarly, Cho and Pappagari [2] applied LSTM with multi-resolution CNN on IEMOCAP, achieving a 21% relative improvement though gains over single acoustic models were modest at 3.4% due to limited generalization. Expanding linguistic diversity, Sultana [3] introduced the SUBESCO Bangla corpus, reporting recognition rates above 70% with anger achieving the highest accuracy (78.3%) and fear the lowest (67.2%), while highlighting issues of ground truth validation and gender bias.

In a different domain, Hajdú-Szücs et al.[4] developed a squash evaluation system combining Gaussian models, TDOA localization, and feed-forward networks, yielding 98% accuracy in controlled conditions but dropping to 88% in live matches where noise and rare events hindered performance. Beyond computational models, Schwartz and Pell [5] explored cross-modal priming of prosody and semantics, showing prosody dominance with up to 99% accuracy, though no practical recognition framework was presented. On the modeling front, Kutlimuratov and Cho [6] proposed a CNN-LSTM with attention that achieved near-perfect accuracy on TESS (99.8%) and strong results on RAVDESS (95.7%), yet its robustness in noisy, real-world contexts remained uncertain.

Advancing hybrid designs, Hasan et al. [7] introduced EmoFormer, a CNN-Transformer model trained on the META EARS dataset, reaching 90% accuracy for five emotion classes but declining to 83% and 70% for larger emotion sets, underscoring scalability challenges. Complementarily, Pan and Wu [8] developed a CLDNN framework integrating 1D CNN, LSTM, and DNN with data augmentation, reporting 91.87% accuracy on RAVDESS and 88.22% on EMO-DB, with noise injection and pitch-shifting shown to enhance generalization though class imbalance persisted. Finally, Ouyang [9] applied CNN-LSTM with MFCC features on a merged SAVEE and RAVDESS dataset, yielding 61.07% accuracy overall, with anger (75.31%) and neutral (71.70%) recognized most effectively, while disgust lagged (38.33%) due to overlapping negative emotions.

Collectively, these studies demonstrate the promise of hybrid deep learning and augmentation strategies in SER, while highlighting persistent limitations in dataset diversity, scalability to many emotion classes, noise robustness, and cross-domain generalization.

3 Proposed Methodology

The proposed methodology for this project will consist of the following steps:

1. Dataset Collection and Preprocessing

- We will use the Toronto Emotional Speech Set (TESS) containing 2,800 utterances (2 seconds each) from two female speakers across seven emotions.
- All audio will be resampled to a common rate (e.g., 16 kHz) and trimmed to remove leading and trailing silence.
- We will apply data augmentation techniques such as noise addition and pitch shifting to increase variability.

2. Feature Extraction

- We will compute Mel-scaled spectrograms or Mel-frequency cepstral coefficients (MFCCs) as 2D time-frequency inputs.
- These image-like representations will be used as inputs to the CNN layers for local acoustic feature extraction.

3. Model Development

- **CNN - LSTM:** We will apply 2D convolution and pooling layers to extract spatial features, reshape feature maps into sequences, and use LSTM layers to model temporal dynamics. A final softmax layer will output emotion probabilities.
- **CNN - Transformer:** We will use the same CNN stack for local feature extraction, followed by Transformer encoder blocks with self-attention to capture long-range dependencies. Outputs will be aggregated using global pooling before classification.

4. Training Procedure

- We will train both models using supervised learning with categorical cross-entropy loss.
- The dataset will be split into training, validation, and test sets (e.g., 80%/10%/10%) or k-fold cross-validation will be used.
- We will use the Adam optimizer with learning-rate scheduling and early stopping.
- Dropout and batch normalization will be applied to improve generalization.

5. Evaluation Plan

- We will evaluate the models using accuracy, precision, recall, and F1-score.
- Confusion matrices will be analyzed to identify commonly confused emotions.
- Identical preprocessing and data splits will be ensured for fair model comparison.

References

[1] S. A. K. Basha, "Exploring deep learning methods for audio speech emotion detection: An ensemble mfccs, cnns and lstm," *Applied Mathematics & Information Sciences*, vol. 19, Jan 2025.

[2] J. Cho and R. Pappagari, "Deep neural networks for emotion recognition combining audio and transcripts," *arXiv preprint*, Nov 2019.

[3] S. Sultana, "Sust bangla emotional speech corpus (subesco): An audio-only emotional speech corpus for bangla," *PLOS ONE*, Apr 2021.

[4] P. Hajdú-Szücs *et al.*, "Audio-based performance evaluation of squash players," *PLOS ONE*, Apr 2020.

[5] S. Schwartz and M. D. Pell, "Emotional speech processing at the intersection of prosody and semantics," *PLOS ONE*, Oct 2012.

[6] B. Kutlimuratov and J. Cho, "Hybrid lstm–attention and cnn model for enhanced speech emotion recognition," *Applied Sciences*, vol. 14, p. 11342, Dec 2024.

[7] R. Hasan, M. Nigar, N. Mamun, and S. Paul, "Emoformer: A text-independent speech emotion recognition using hybrid transformer-cnn model," in *Proc. 2024 27th International Conference on Computer and Information Technology (ICCIT)*, (Cox's Bazar, Bangladesh), Dec 2024.

[8] 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, p. 2436, May 2023.

[9] Q. Ouyang, "Speech emotion detection based on mfcc and cnn-lstm architecture," *arXiv preprint*, 2025.