IARCO_2025 - Apurbo Kumar.pdf

by Sanaul Haque

Submission date: 13-Oct-2025 10:12PM (UTC+0700)
Submission ID: 2779906419
File name: IARCO_2025_-_Apurbo_Kumar.pdf (604.69K)
Word count: 1112
Character count: 6697

IARCO RESEARCH PROPOSAL

on

Emotion Recognition in Human Speech Using Deep Learning Approaches

Scholar's Name: Apurbo Kumar

Institution: Bangladesh University of Engineering and Techonology

Category: Senior

Year: 4th Year, BSc.

Country: Bangladesh

Major: Civil Engineering

Submission Date: 28/09/2025

Registered Email Address: apurbokumar1355@gmail.com

Research Problem

Human communication is inseparable from emotions, which influence not only the words we speak but also how we say them. Emotion recognition has therefore become an essential challenge for artificial intelligence systems that aim to provide natural, human like interaction [4], [6]. Speech is one of the most powerful carriers of emotion, containing rich information embedded in tone, pitch, rhythm, and intensity. However, most existing AI systems still focus on the linguistic message and fail to capture emotional cues [4].

Traditional Speech Emotion Recognition (SER) methods have relied on hand-engineered acoustic features such as prosody and spectral properties, paired with machine learning classifiers like Support Vector Machines (SVMs) or Gaussian Mixture Models (GMMs). These methods demonstrated modest success but lacked robustness across speakers, languages, and noisy real-world conditions [4], [6].

Deep learning has revolutionized SER by automatically learning features from data. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and hybrid CNN–RNN architectures have achieved promising results [2], [3], [5]. Despite this progress, key challenges remain unresolved: confusion between acoustically similar emotions such as calm vs. neutral or fear vs. surprise; reliance on acted datasets rather than natural speech; and limited cross-cultural adaptability. This research aims to address these gaps by systematically comparing CNNs, RNNs, and hybrid CNN–RNNs using the RAVDESS dataset, with the objective of determining which architecture best balances accuracy, generalization, and robustness [1].

Existing Literature

Research in SER has evolved through three broad stages: traditional feature-based approaches, the introduction of perceptual features, and the deep learning era [4], [6].

1. Traditional Approaches: Earlier studies used prosodic features (pitch, energy, rhythm) and spectral features (formants, cepstral coefficients) combined with statistical classifiers like SVMs, KNN, and GMMs. While these approaches performed reasonably well in small datasets, they often failed when generalized to larger and more diverse speech samples_[4],[6].

2. Perceptual Features: The development of perceptually relevant features like Mel-

Frequency Cepstral Coefficients (MFCCs) improved SER performance. MFCCs capture human auditory perception more effectively, enabling better differentiation between emotions. However, feature extraction remained heavily manual, limiting scalability.

3. Deep Learning Advances: Deep learning introduced automatic feature extraction. CNNs treat spectrograms as images and excel at identifying spatial time–frequency patterns [2], [5], while RNNs (especially LSTMs and GRUs) capture sequential dependencies in audio signals [3]. Hybrid CNN–RNN models combine these strengths, leading to state-of-the-art accuracy on benchmark datasets [2], [5].

4. Key Datasets: Datasets have played a crucial role in advancing SER. EMO-DB provided early benchmarks but is small in size. CREMA-D increased diversity by including multiple actors and expressions. RAVDESS stands out for its balance, quality, and availability of both speech and song recordings, making it a strong candidate for deep learning studies [1].

Table 1: Summary of SER Studies

Author/Year	Dataset	Model	Accuracy Reported	Key Insights
Fayek et al. (2017)	EMO-DB	CNN	~75%	CNNs capture localized acoustic patterns
Latif et al. (2020)	CREMA-D	LSTM/GRU	~78%	RNNs model temporal dependencies in speech
Tzirakis et al. (2018)	RAVDESS	End-to-end DNN	~81%	End-to-end learning reduces manual features
Recent Hybrid Models	RAVDESS	CNN–RNN	85–90%	Combining spatial + temporal features works best

Research Question

To what extent can Deep Learning (DL) architectures like CNNs, RNNs, and hybrid CNN–RNN models improve the accuracy of speech emotion recognition, mainly in distinguishing acoustically similar emotions like neutral vs. calm and fear vs. surprise?

Methodology

This study adopts an experimental design to systematically evaluate CNNs, RNNs, and CNN–RNN hybrids.

Dataset: The RAVDESS dataset will be used, containing recordings of eight emotions neutral, calm, happy, sad, angry, fearful, surprise, and disgust performed by 24 actors. It offers balanced, high-quality samples suitable for deep learning research.

Preprocessing: Convert audio to mono, normalize volume, and resample for consistency. Extract features: MFCCs and Mel-spectrograms. Apply data augmentation (pitch shifting, time stretching, noise addition) to improve generalization [4], [6].

Models:

- CNN Model: Focus on spectrograms to capture spatial features [2], [5].
- RNN Model: Apply LSTM/GRU layers to MFCC sequences for temporal dependencies [3].
- Hybrid CNN–RNN: Use CNN layers for local feature extraction, followed by RNN layers for sequence modeling [2], [3].

Training & Evaluation: Models will be trained using the Adam optimizer with categorical cross-entropy loss. Metrics include accuracy, precision, recall, F1-score, and confusion matrices. Cross-validation will be used to ensure robust performance.

Tools: Implementation will use Python, TensorFlow/Keras, librosa, and scikit-learn.

Research Topic

Emotion recognition in speech is increasingly important as AI systems expand into sensitive, human-centered domains. In healthcare, SER can assist in stress monitoring and depression screening. In education, it can enable adaptive learning platforms that respond to students' needs. In customer service, emotion-aware chatbots and assistants can provide more empathetic interactions. In robotics, SER enables more natural human and machine collaboration [3].

This research contributes by providing a detailed comparative analysis of CNN, RNN, and hybrid CNN–RNN models on SER. Findings will guide future development of emotionally intelligent AI systems that go beyond functional performance and incorporate human-centered responsiveness [2], [5].

References

[1] S. R. Livingstone and F. A. Russo, “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS),” PloS One, vol. 13, no. 5, p. e0196391, 2018.

[2] H. M. Fayek, M. Lech, and L. Cavedon, “Evaluating deep learning architectures for speech emotion recognition,” Neural Networks, vol. 92, pp. 60–68, 2017.

[3] S. Latif, R. Rana, J. Qadir, and J. Epps, “Deep representation learning in speech processing: Challenges, recent advances, and future trends,” Computer Speech & Language, vol. 59, p. 101119, 2020.

[4] S. G. Koolagudi and K. S. Rao, “Emotion recognition from speech: A review,” Int. J. Speech Technol., vol. 15, no. 2, pp. 99–117, 2012.

[5] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, “End-to-end speech emotion recognition using deep neural networks,” in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process. (ICASSP), 2018, pp. 5089–5093.

[6] D. Ververidis and C. Kotropoulos, “Emotional speech recognition: Resources, features, and methods,” Speech Communication, vol. 48, no. 9, pp. 1162–1181, 2006.

IARCO_2025 - Apurbo Kumar.pdf

ORIGINALITY REPORT

17%	14%	15%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	arxiv.org Internet Source	4%
2	par.nsf.gov Internet Source	2%
3	impa.usc.edu Internet Source	2%
4	gyan.iitg.ac.in Internet Source	1%
5	www.researchgate.net Internet Source	1%
6	doaj.org Internet Source	1%
7	deepai.org Internet Source	1%
8	aclanthology.org Internet Source	1%
9	Aparna Vyakaranam, Tomas Maul, Bavani Ramayah. "Comparison of three hybrid architectures using 1D, 2D, and 3D CNNs for speech emotion recognition", International Journal of Speech Technology, 2025 Publication	1%
10	Arvind Dagur, Karan Singh, Pawan Singh Mehra, Dhirendra Kumar Shukla. "Intelligent	1%

Computing and Communication Techniques - Volume 3", CRC Press, 2025
Publication

11	S.P. Jani, M. Adam Khan. "Applications of AI in Smart Technologies and Manufacturing", CRC Press, 2025 Publication	1 %
12	Arziki Pratama, Sari Widya Sihwi. "Speech Emotion Recognition Model using Support Vector Machine Through MFCC Audio Feature", 2022 14th International Conference on Information Technology and Electrical Engineering (ICITEE), 2022 Publication	1 %
13	Abu Quwsar Ohi, M. F. Mridha, Md. Abdul Hamid, Muhammad Mostafa Monowar. "Deep Speaker Recognition: Process, Progress, and Challenges", IEEE Access, 2021 Publication	1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	Off		