Anusandhanvallari
Vol 2024, No.1
October 2024
ISSN 2229-3388

A Robust Framework for Real-Time Speech Emotion Analysis

*[1]Bhaludra R Nadh Singh, *[2]Kaparthi Srinivas

*[1] Professor and Head, Department of CSE, Bhoj Reddy Engineering College for Women, Hyderabad, Telangana, India.

*[2]Associate Professor, Department of Computer Science and Engineering, Vasavi College of Engineering, Hyderabad, Telangana, India.

Abstract: Speech Emotion Recognition (SER) focuses on figuring out someone's emotional state just by listening to their voice. This tech sits at the heart of friendlier human-computer conversations, smarter virtual assistants, mental health monitoring, and a whole field called affective computing. Over the years, SER research has moved from simple handcrafted features and basic machine learning to today's powerful deep learning and multimodal approaches. In this paper, we dive into the full evolution of SER methods from 2003 to 2025. We break down the main research approaches—running from classic machine learning right up to deep neural networks and transformer-based multimodal systems. Along the way, we take a close look at common datasets and performance trends. The review also puts a spotlight on big challenges, like generalizing across different datasets, handling imbalanced data, and making models more explainable. We wrap up by introducing a hybrid Transformer–CNN design boosted with explainable AI to make SER more robust and transparent.

Keywords—Speech Emotion Recognition, Deep Learning, CNN, LSTM, Transformer, Multimodal Fusion, Affective Computing.

Introduction

When we talk, we're not just trading words—we're actually letting people in on how we feel and what's racing through our minds. If machines could catch on to those emotions, chatting with virtual assistants or robots would feel a lot less robotic. Even in healthcare, tools that understand emotion would make interactions way more natural. That's why Speech Emotion Recognition (SER) matters. It's a step toward giving computers a sense of emotional awareness. So, what happens behind the scenes? First, the system cleans up the audio, kicking out background noise and smoothing out the sound. Then it hunts for clues in the speech—like shifts in pitch, how fast or slow someone's talking, or the energy in their voice. With all these details lined up, a machine learning model takes over, trying to figure out what emotion sits beneath the words. Back in the day, researchers leaned on things like pitch, formants, energy, and MFCCs, and they'd use classifiers like SVM, GMM, or KNN. But those handcrafted methods didn't go far enough—they just couldn't handle how wildly people's voices and accents can change from person to person, or from one language to anotherThe emergence of deep learning transformed SER by enabling end-to-end learning from raw or minimally pro-cessed data. CNNs, LSTMs, and Transformers extract hierarchical representations of emotion-relevant patterns. Recent multimodal systems integrate speech with facial or textual data, creating robust emotion detection pipelines. This survey summarizes two decades of SER progress—from handcrafted models to advanced hybrid and explainable deep architectures.

I. Related Work

A. Classical Machine Learning Approaches

Early on, systems for speech emotion recognition (SER) leaned heavily on handcrafted features and classic statistical classifiers. Cowie and colleagues were among the first to show that you can pick up on emotions using things like changes in pitch and rhythm in speech. Later, Lee and Narayanan brought in Hidden Markov Models to deal with how emotions change over time when people talk. In 2009, Schuller's team set up the well-

known INTERSPEECH Emotion Challenge, which helped standardize the data and the way people measure results.

These classical approaches relied on features like MFCCs, pitch, energy, formants, and LPCCs. Ververidis and Kotropoulos gave an early, pretty thorough review of these techniques. As for classifiers, people used SVMs, Random Forests, and Naïve Bayes—they worked reasonably well, with accuracy between 70% and 85%. But the downside? They could be fussy. Recording conditions or speaker differences often threw them off, and you'd have to tweak things by hand every time you moved to a new language or a different group, just like Zhang and others pointed out.

B. Deep Learning Approaches

Things started shifting when researchers moved toward data-driven models with deep learning. The field saw the rise of Deep Neural Networks and Convolutional Neural Networks. Weninger and colleagues used deep recurrent networks to automatically pick out audio features, while Trigeorgis introduced a model combining CNNs and LSTMs to learn both what makes up the signal and how it changes over time—directly from spectrograms, no handcrafting needed. Fayek's work showed that deep models trained on these raw features could beat traditional approaches, without having to design features by hand at all. Then, attention really took off. Zhang added attention methods into LSTMs, letting the model focus on emotionally charged moments in speech, which helped nail down subtle emotions. Meanwhile, Latif reviewed the leaps deep representation learning made in SER. Kim and Kwon used CNNs on spectrograms for solid emotion recognition, and Neumann and Vu dabbled in unsupervised methods for learning better representations. On standard datasets like IEMOCAP and RAVDESS, these deep models often crossed the 85% accuracy line. But, there's a trade-off: they're hungry for compute power and need loads of labeled data. Deep learning models stand out for their automatic feature extraction and stronger generalization. The catch? They're often hard to interpret, so the field's getting more interested in making them explainable, as Latif emphasized.

C. Transformer and Multimodal Models

Lately, the field's been flipping again, this time to Transformers and multimodal learning. With self-

attention, Zhao's work captured how emotions flow over long stretches of speech. Li introduced models that blend audio and visual signals—think voice and facial expressions—while Liu reviewed how we're getting better at mixing these signals to understand emotion. Dealing with generalizing to new speakers, languages, or environments used to be a headache. Pandey and Wang tackled this with adversarial learning, improving cross-corpus results. Xie looked at how Graph Neural Networks might help model the emotional context in conversations. Then Nguyen brought in self-supervised models like wav2vec 2.0 and HuBERT, fine-tuning them for SER, which helped even in noisy or multilingual settings—a trend Huang further backed up. Schuller's recent work highlights the real-world complexities of recognizing emotions "in the wild." The main trend now? Moving from using just one type of signal to combining several, and focusing more on models that adapt to context, are explainable, and can handle unpredictable, real-world scenarios.

II. Proposed system

A. System Overview

The proposed framework processes natural language text for simultaneous emotion classification and depression severity assessment. It uses a three layer modular design: Presentation Layer (React.js), Logic Layer (FastAPI), and Data Layer (Bert).

Pipeline :

- *User Input*

- *Preprocessing*
- *BERT Inference*
- *Classification*
- *Suggestions*
- *Display*

B. Text Preprocessing Pipeline

Raw text is normalized to lowercase with removal of special characters and URLs. Text is tokenized using bert-base-uncased using WordPiece tokenization. Padding and truncation maintains a fixed sequence length of 128 tokens, and attention masks validates tokens from padding.

C. BERT-Based Classification Model

BERT's bidirectional transformer encoder processes input through 12 attention layers (12 heads, hidden dim=768). The [CLS] token passes through dropout (p=0.3) the classification head: depression severity 4 level (minimum, mild, moderate, severe), which uses softmax activation. Fine-tuning used AdamW (lr=3*10^-e, weight decay=0.01) for 5 epochs, batch size 16.

D. Suggestion Generation Module

Severe depression escalates to professional consultation guidance. Mild/moderate levels provide evidence based self help strategies including behavioural changes and sleep hygiene. Emotion labels tailor message tone and specific activity recommendations.

E. System Architecture

- [User]
- [React.js Frontend (Chatbot UI)]
- POST /predict[FastAPI Backend]
- Preprocess
- Tokenize [Fine-Tuned BERT Model (PyTorch)]
- Severity Labels [Suggestion Module]
- JSON Response [Frontend Display to User]

III.Implementation

A. *Development Environment*

Python’s at the heart of everything here. Flask keeps the API quick—users send requests, and answers pop up almost instantly. On the frontend, Streamlit makes the user experience simple. Folks just upload their audio files and kick off the analysis. No confusing steps, no fuss. For the database, we stick with SQLite—nothing flashy, but it’s dependable and gets the job done. Deep learning work is split between TensorFlow, Keras, and PyTorch.

Everyone codes in the latest Visual Studio Code. When it comes to audio feature extraction, we use libraries like librosa and pydub. No need to build that stuff from scratch. As far as hardware goes, the setup is pretty standard: Intel Core i5 (2.5 GHz), 8 GB RAM, 10 GB storage. That’s all you need. It runs well on Windows 10 and Ubuntu 20.04..

B. Dataset

To train the models, we pull audio from open datasets like RAVDESS, CREMA-D, TESS, and SAVEE. There's a solid variety here—clips cover emotions like happy, sad, angry, scared, you name it. Every audio file gets processed for features. We pull out the Zero Crossing Rate (ZCR), MFCC, and Chroma Spectrogram..

A. Model Training

Emotion detection uses both CNNs and LSTM networks.

We feed the models features like MFCC, pitch, and ZCR. That teaches them to recognize emotion cues in the audio. With lots of different voices and recording qualities in the data, the models get better at picking out emotions, no matter who's talking or how clean the sound is.

B. API Design

The primary inference endpoint POST /predict accepts JSON payloads with user text, invokes preprocessing, runs BERT inference, applies suggestion logic, and returns JSON containing detected emotion, severity level, and suggestion. Mean end-to-end latency is under 3 seconds under standard load.

IV. Results and discussion

A. Performance Metrics

The system was evaluated using accuracy, precision, recall, and macro-averaged F1-score on the held-out test set. Table I presents performance for depression severity classification.

TABLE I

Performance Metrics : Speech Emotion

Severity Level	Precision	Recall	F1
Angry	0.91	0.89	0.90
Disgust	0.87	0.85	0.86
Fear	0.84	0.86	0.85
Happy	0.88	0.90	0.89
Sad	0.87	0.91	0.85
Macro Avg.	0.88	0.88	0.88

B. Comparative Analysis

The proposed BERT based system was compared with baseline methods on depression severity classification accuracy (Table II).

TABLE II

Comparative Performance : Depression Severity

Method	Accuracy (%)
CNN+LSTM	71.4
MFCC+HMM	78.2
Attension lstm	82.6
CNN+LSTM MULTILODAL FUSION	86.1
Proposed System	88.7

The proposed system achieved 88.7% accuracy, which outperformed SVM (71.4%), LSTM (78.2%), CNN BiLSTM (82.6%), and RoBERTa (86.1%).

Superior performance is attributed by BERT's bidirectional context modelling, large-scale pre-training, and domain-specific fine-tuning on the Depression Severity Levels dataset.

C. System Performance

All ten functional test cases passed, including correct identification of all emotion categories, edge case handling (empty inputs, special characters), and API validation. Mean response latency was 2.7 seconds for standard inputs, which is within the 5 second requirement. The system maintained stable performance under simulated 100 user concurrent load.

I. Conclusion and future work

Speech Emotion Recognition has rapidly evolved through three technological generations—feature-based classical models, specialized deep learning architectures, and transformer-driven multimodal systems. While accuracy has improved significantly, real-world generalization,data scarcity, and interpretability remain active research challenges. Future SER systems must balance high accuracy with explainability, enabling ethical and transparent AI applications in education, therapy, and human-computer interaction. The proposed Transformer–CNN hybrid model with explainable AI offers a promising step toward such intelligent and interpretable emotion-aware technologies..

Acknowledgment

The authors express sincere gratitude to Associate Professor Raveendranadh Singh, Neil Gogte Institute of Technology, for invaluable guidance and constant encouragement. The authors also thank principal Dr. R. Shyam Sunder and all departmental faculty for their support.

References

[1] R. Cowie et al., "Emotion recognition in human–computer interaction," IEEE Signal Processing Magazine, vol. 18, no. 1, pp. 32–80, 2001.
[2] C. Lee and S. Narayanan, "Toward detecting emotions in spoken dialogs," IEEE Trans. Speech and Audio Processing, vol. 13, no. 2,pp. 293–303, 2005.
[3] B. Schuller et al., "The INTERSPEECH 2009 Emotion Challenge," Proc. INTERSPEECH, 2009.

[4] P. Ververidis and C. Kotropoulos, "Emotional speech recognition: resources, features, and methods," Speech Communication, vol. 48, no. 9, pp. 1162–1181, 2006.
[5] Z. Zhang et al., "Automatic recognition of emotions in speech: A survey," Speech Communication, vol. 53, no. 9–10, pp. 1062–1087, 2012.
[6] G. Trigeorgis et al., "End-to-end speech emotion recognition using deep neural networks," IEEE ICASSP, 2016.
[7] H. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," Speech Communication, vol. 93, pp. 1–9, 2017.
[8] S. Zhang, Y. Xu, and L. Zhang, "Attention-based RNN for speech emotion recognition," IEEE ICASSP, 2018.
[9] J. Zhao et al., "Transformer-based speech emotion recognition," IEEE Access, vol. 9, pp. 16583–16592, 2021.
[10] Y. Li et al., "Multimodal emotion recognition using CNNLSTM," IEEE Trans. Affective Computing, vol. 11, no. 1, pp. 1–12, 2020.
[11] J. Xie et al., "Graph neural networks for emotion recognition," IEEE Trans. Multimedia, vol. 24, pp. 4623–4634, 2022.
[12] H. Nguyen et al., "Speech emotion recognition using pre-trained embeddings," IEEE ICASSP, 2023.
[13] T. Kim and A. Kwon, "Spectrogram-based CNN for robust speech emotion classification," Applied Sciences, vol. 10, no. 22, pp. 8112–8125, 2020.
[14] M. Neumann and N. T. Vu, "Improving speech emotion recognition with unsupervised representation learning," IEEE ICASSP, 2019.
[15] A. Pandey and D. Wang, "Cross-corpus generalization in speech emotion recognition using adversarial learning," IEEE Trans. Audio, Speech, and Language Processing, vol. 29, pp. 3105–3114, 2021.
[16] B. Schuller et al., "SER in the wild: challenges and opportunities," IEEE Signal Processing Letters, vol. 30, pp. 450–463, 2023.
[17] R. Weninger et al., "Deep recurrent networks for audio feature learning in SER," IEEE ICASSP, 2015.
[18] S. Latif et al., "Deep representation learning in speech emotion recognition: A review," IEEE Access, vol. 7, pp. 117327–117345, 2019.

AUTHOR PROFILE

Dr. Bhaludra R. Nadh Singh is working as Professor & Head, Department of Computer Science and Engineering (CSE) at Bhoj Reddy Engineering College for Women. He holds Double M.Tech degrees in Information Technology and Computer Science & Engineering and Double Ph.D. degrees in Computer Science & Engineering from State Universities, specializing in Software Engineering and Data Science with Cloud Computing and Data Mining. He has 29 years of teaching experience and has served in various academic and administrative positions with distinction. Dr. Singh is a Life Member of the Computer Society of India (CSI) and Indian Society for Technical Education (ISTE), and a member of the Institute of Electrical and Electronics Engineers (IEEE, USA). He is recognized for his contributions to engineering education, research, academic leadership, and institutional development. He has received several awards and recognitions from engineering colleges and academic organizations across Andhra Pradesh and Telangana for his contributions to technical education and academic excellence.