JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A review on Speech Emotion Recognition from Raw Audio using LSTM and Neural Networks

Sanjay Chilhate[1], Anjana Verma[2], Nitya Khare[3]
Sagar Institute of Research and Technology - Excellence[1,2,3]

ABSTRACT

Speech Emotion Recognition (SER) has emerged as a crucial domain within human-computer interaction (HCI), enabling machines to identify and respond to users' emotional states. Unlike traditional text-based sentiment analysis, SER relies on auditory cues, making it more complex due to the dynamic and nuanced nature of speech. With the proliferation of deep learning, especially architectures like Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNNs), significant strides have been made in extracting emotional patterns from raw audio data. This review paper delves into recent advancements in SER, focusing primarily on methodologies that bypass extensive feature engineering by utilizing raw waveform data. We comprehensively analyze ten state-of-the-art studies that have contributed novel techniques, including attention mechanisms, hybrid CNN-LSTM models, and end-to-end learning paradigms. Each method is evaluated based on the dataset used, performance metrics (accuracy, F1-score, etc.), and its limitations. A key insight from the review is the increasing reliance on raw audio inputs, eliminating the dependency on handcrafted features such as MFCC or spectrograms. However, existing approaches still struggle with generalization, dataset imbalance, and speaker variability. The paper also presents a brief overview of a proposed LSTM-based architecture designed to enhance robustness across diverse speech signals. Our findings highlight the gaps in current research and suggest directions for future exploration, particularly emphasizing multilingual datasets and unsupervised learning techniques for SER.

Keywords
Speech Emotion Recognition (SER), Raw Audio, LSTM, Deep Learning, CNN, Attention Mechanism, Human-Computer Interaction, Emotion Detection, End-to-End Learning, Neural Networks

1. INTRODUCTION

Speech is a primary mode of human communication, not only conveying information but also reflecting a speaker's emotional state. Recognizing and interpreting emotions from speech has become increasingly important in fields such as healthcare, virtual assistants, customer service, and affective computing. This field, widely known as Speech Emotion Recognition (SER), enables machines to understand human emotions, contributing significantly to the development of emotionally intelligent systems and enhancing user experience in Human-Computer Interaction (HCI).

Traditionally, SER systems relied heavily on hand-engineered features such as Mel-Frequency Cepstral Coefficients (MFCCs), Chroma, Spectral Contrast, and Zero-Crossing Rate, which required significant domain expertise and often failed to generalize across languages and speaker conditions. With the advent of deep learning, the research paradigm has shifted from manual feature extraction to end-to-end learning, where systems learn directly from raw audio waveforms. Among various deep learning architectures, Long Short-Term Memory (LSTM) networks have demonstrated significant promise due to their ability to capture temporal dynamics in sequential data like speech.Recent literature also explores hybrid models such as CNN-LSTM and attention-based mechanisms that combine local feature extraction and sequential learning capabilities. The focus of this review is specifically on models that use raw audio input, minimizing pre-processing and eliminating the dependence on domain-specific audio features. This approach aligns with the broader trend in deep learning toward fully automated learning systems. Despite notable progress, several challenges persist: emotion datasets often lack diversity in language and speaker variability; real-world emotion expressions can be ambiguous and context-dependent; and deep models often demand large amounts of labeled data. The goal of this review paper is to summarize existing research efforts addressing these challenges and highlight the gap that motivates the proposed methodology—an LSTM-based model trained directly on raw audio data.

2. BACKGROUND AND MOTIVATION

Speech Emotion Recognition (SER) is a subfield of affective computing that aims to identify emotional states such as happiness, anger, sadness, or fear from audio signals. Unlike textual sentiment analysis, SER processes acoustic features like pitch, tone, rhythm,

and energy, which are often more expressive of emotions. Early systems primarily depended on engineered acoustic features—such as MFCCs, Linear Predictive Coding (LPC), and formant frequencies—combined with classical machine learning algorithms like Support Vector Machines (SVMs) or Hidden Markov Models (HMMs). Although these methods achieved moderate success, they suffered from limitations related to feature selection **BIAS** and generalization across unseen speakers or languages. With the rise of deep learning, particularly Recurrent Neural Networks (RNNs) and their variant Long Short-Term Memory (LSTM) networks, SER research has experienced a significant shift. LSTM is particularly well-suited for sequence modeling, allowing the network to learn long-range temporal dependencies in speech, which are crucial for emotion classification. In addition, Convolutional Neural Networks (CNNs) have been widely adopted to capture local patterns in audio signals, especially when transformed into spectrograms or mel-scale representations.

However, converting raw audio into feature representations like spectrograms introduces information loss and adds extra computational complexity. This challenge has encouraged researchers to explore models that learn directly from raw waveform audio, skipping traditional preprocessing stages. Such models, often trained end-to-end using deep neural networks, have shown potential in maintaining the richness of emotional cues embedded in speech.

The motivation behind this review stems from several emerging trends:

1. The need for generalizable SER models that work across different languages, genders, and speaking styles.
2. The growing availability of emotional speech datasets such as RAVDESS, TESS, CREMA-D, and EmoDB.
3. The increasing adoption of LSTM and hybrid CNN-LSTM architectures in end-to-end SER systems.
4. The shift toward raw-audio-based input to reduce manual feature engineering and ensure greater flexibility in real-world deployment.

By systematically reviewing state-of-the-art approaches that utilize LSTM and other neural network models for SER from raw audio, this paper aims to uncover the current research landscape, identify key gaps, and motivate the proposed methodology for improved performance and robustness.

3. LITERATURE REVIEW

In this section, we review ten significant research studies in the field of Speech Emotion Recognition (SER), with a focus on models utilizing deep learning architectures like LSTM and raw audio-based inputs. Each study is summarized with key methods used, performance results, and identified gaps.

[1] Yang, Y., Zhang, L., & Liu, C. (2024)

Yang et al. developed a dual-layer LSTM model to improve long-term emotional pattern recognition in raw speech. The model achieved a 2% improvement over standard LSTM baselines on the RAVDESS dataset. However, it was evaluated on a limited number of datasets and lacked robustness testing across languages and accents.

[2] Zhang, Q., Chen, Y., & Wang, H. (2024)

This study implemented SincNet layers followed by LSTM for raw waveform emotion recognition. Tested on IEMOCAP, it achieved 85.1% accuracy. While innovative, its over-reliance on IEMOCAP affects generalizability to multilingual or noisy environments.

[3] Mou, L., Zhang, H., & Ghosh, S. (2021)

Mou and co-authors proposed a CNN-LSTM fusion for speech emotion tasks. Their system exceeded traditional models in both accuracy and training stability. However, it used spectrogram-based input rather than raw waveforms, making it less suitable for real-time low-resource systems.

[4] Ahmed, R., Bhattacharya, S., Kumari, A., & Patel, M. (2021) ***(Indian contribution)***

This Indian-led research employed a hybrid 1D-CNN + LSTM + GRU architecture, enriched with data augmentation. It was tested on RAVDESS, EMO-DB, TESS, and SAVEE, achieving accuracies above 90%. Despite strong performance, it depended heavily on handcrafted features and lacked a raw waveform pipeline.

[5] Kilimci, Z. H., Anbarjafari, G., & Singh, A. (2023) ***(Indian co-author)***

The team evaluated CNNs on raw waveforms for TESS and RAVDESS, achieving 95.86% accuracy. While excellent results were reported, the paper lacked comparative benchmarking with LSTM or BiLSTM-based networks.

[6] Patel, R., Kaur, J., Sharma, P., & Sahu, R. (2022) ***(Indian contribution)***

The authors developed a BiLSTM-based model that leveraged log-mel spectrograms and MFCCs for SER. Trained on the CREMA-D and RAVDESS datasets, their system performed well but required heavy pre-processing. No end-to-end raw audio learning was implemented.

[7] Diwan, S., Aggarwal, A., & Rane, V. (2023) ***(Indian contribution)***

This team built a transformer-LSTM hybrid for SER, using raw audio input with adaptive spectrogram transformation layers. The model achieved notable robustness in noisy conditions. However, training required significant compute resources, limiting real-time applications.

[8] Anvarjon, A., & Kwon, O. J. (2020)

They introduced a CNN-based system that emphasized deep frequency features for lightweight SER deployment. While computationally efficient, it lacked LSTM integration and failed to leverage temporal speech dynamics fully.

[9] Sharma, S., Prasad, P., & Iyer, M. (2021) ***(Indian contribution)***

The authors proposed an attention-enhanced Bi-LSTM model for SER, tested on multilingual datasets including Hindi and Tamil emotional speech corpora. Their method handled tonal variations effectively but lacked raw audio end-to-end learning.

[10] Wang, T., Yu, L., & Joshi, M. (2020) ***(Indian co-author)***

This team developed a DS-LSTM model using dual-inputs of MFCC and mel-spectrogram features. The model showed improvements over single-stream LSTMs with 72.7% weighted accuracy. However, raw waveform inputs were not explored.

Summary of Research Gaps

- **Underrepresentation of Raw Audio Approaches:** Most Indian and global studies rely on MFCCs or spectrograms. Direct processing of raw waveforms remains limited.
- **Temporal Modeling Weakness:** Although CNNs dominate the field, fewer studies integrate robust temporal models like BiLSTM or attention-enhanced LSTM with raw audio inputs.
- **Dataset Diversity:** Limited use of Indian language datasets or mixed-lingual speech corpora. Most studies test on English-only datasets like RAVDESS or IEMOCAP.
- **Real-time Constraints:** Despite model innovations, many are computationally expensive, limiting mobile or real-time deployments.
- **End-to-End Learning:** A common gap remains in developing fully end-to-end pipelines that start from raw waveforms and perform emotion recognition without handcrafted preprocessing.

4. PROPOSED METHOD

Building upon the gaps identified in the literature, we propose an **end-to-end speech emotion recognition (SER) system** that directly operates on **raw audio waveforms**. The system leverages the **temporal strength of LSTM** networks combined with the **feature learning capability of 1D Convolutional Neural Networks (CNNs)**. This hybrid approach allows the model to extract relevant emotional patterns without relying on handcrafted features like MFCCs or spectrograms, which often lead to information loss.

Key Characteristics of the Proposed System:

- **Raw Audio Input**: Instead of converting the signal into MFCC or mel-spectrograms, we directly feed normalized raw waveform samples to the model.
- **1D Convolution Layer**: Acts as a learnable front-end to extract frequency- and time-local patterns from the waveform. It also reduces the dimensionality of raw signals, similar to the role of SincNet or Sinc-Convolutions but without handcrafted filters.
- **Stacked LSTM Layers**: Two LSTM layers are used to capture long-term temporal dependencies in emotional speech. These layers model the progression and evolution of speech emotion over time.
- **Attention Mechanism (optional)**: An attention layer is optionally included to focus the model's learning on emotionally salient parts of the audio, improving classification accuracy.
- **Dense Output Layer**: The final fully connected layer uses Softmax activation to classify emotions such as *happy, sad, angry, fearful, calm,* etc., depending on the dataset.
- **Training on Multiple Datasets**: To ensure generalizability and multilingual robustness, the model will be trained on diverse datasets like **RAVDESS**, **TESS**, **CREMA-D**, and **IEMOCAP**.

Advantages of the Proposed Method:

- **Truly end-to-end training** with no need for feature engineering.
- More **robust to noisy data** due to direct modeling of the waveform.
- Compatible with **real-time inference** using lightweight deployment.
- Flexible across **languages and dialects**, especially useful for Indian multilingual environments.

This methodology will be fully detailed in the **main research paper**, along with flowcharts, algorithmic steps, and performance comparisons against state-of-the-art models.

5. CONCLUSION

This review highlights the evolution and current trends in speech emotion recognition (SER), focusing especially on the growing interest in deep learning approaches using raw audio inputs. The transition from handcrafted feature-based methods (like MFCCs and spectrograms) to end-to-end deep neural architectures marks a significant milestone in SER research. Through an in-depth examination of ten influential studies—both global and Indian—this review reveals key limitations in existing works, including limited generalization, lack of raw waveform processing, insufficient modeling of temporal dependencies, and absence of multilingual dataset testing.

The majority of existing approaches rely on preprocessed acoustic features, which often obscure vital emotional cues embedded in the original waveform. Moreover, models that process speech signals directly, such as those using CNNs or SincNet, frequently neglect long-range temporal dependencies that are essential for understanding emotional tone and rhythm. Our proposed hybrid architecture addresses these issues through a combination of 1D CNN layers for raw feature extraction and LSTM layers for temporal modeling, with an optional attention mechanism to further refine performance

6. REFERENCES

1. Yang, Y., Zhang, L., & Liu, C. (2024). Dual-layer LSTM for Real-Time Emotion Recognition. arXiv preprint arXiv:2411.09189.
2. Zhang, Q., Chen, Y., & Wang, H. (2024). End-to-End Speech Emotion Recognition using SincNet and LSTM. arXiv preprint arXiv:2402.11954.
3. Mou, L., Zhang, H., & Ghosh, S. (2021). Combining CNN and LSTM for Speech-Based Emotion Detection. Proceedings of ROCLING 2021.
4. Ahmed, R., Bhattacharya, S., Kumari, A., & Patel, M. (2021). Hybrid Deep Networks for Multilingual SER. arXiv preprint arXiv:2112.05666.
5. Kilimci, Z. H., Anbarjafari, G., & Singh, A. (2023). Raw Waveform Based Deep Learning for Emotion Recognition. arXiv preprint arXiv:2307.02820.

6. Patel, R., Kaur, J., Sharma, P., & Sahu, R. (2022). Deep Learning Approaches for SER in Indian Languages. Journal of Intelligent Systems, 32(3), 145–160.
7. Diwan, S., Aggarwal, A., & Rane, V. (2023). Transformer-LSTM Based Emotion Recognition from Raw Audio. International Journal of Speech Technology.
8. Anvarjon, A., & Kwon, O. J. (2020). Lightweight CNNs for Audio-Based Emotion Detection. Applied Acoustics, 163, 107224.
9. Sharma, S., Prasad, P., & Iyer, M. (2021). Multilingual BiLSTM Attention Model for SER. IEEE India Conference (INDICON).
10. Wang, T., Yu, L., & Joshi, M. (2020). Dual Sequence LSTM for Audio Emotion Recognition. ICASSP Proceedings 2020.