INTERNATIONAL JOURNAL OF INFORMATIVE & FUTURISTIC RESEARCH (IJIFR)
ISSN: 2347-1697
VOLUME 13||ISSUE 9|| MAY 2026
An Open Access, Peer Reviewed, Cross Reference DOI & Google Scholar Indexed International Research Journal
UGC Approved || IJIFR Impact Factor= 8.057
Estd Year: 2013 || Email: editors@ijifr.org
Availaible Online at: https://www.ijifr.org/

DOI: https://doi.org/10.64672/IJIFR/26.05.13.09.035 PUBLISHED ON: MAY 22, 2026

SPEECH EMOTION RECOGNITION USING DEEP NEURAL NETWORKS: DESIGN AND IMPLEMENTATION OF EMOVOICE AI

Akkala Reddy Charan [1], S. Usharani [2]
[1] Student, [2] Professor & Head,
Department of Computer Applications, Viswam Engineering College,
Andhra Pradesh, India

ABSTRACT

This paper presents EmoVoice AI, a deep learning–based speech emotion recognition system that classifies human emotional states directly from raw audio signals. The proposed framework adopts a hybrid Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) architecture, jointly capturing localized spectral patterns and long-range temporal dependencies. Acoustic features are derived using Librosa as forty Mel-Frequency Cepstral Coefficients (MFCCs) augmented with first- and second-order delta coefficients, yielding a 120-dimensional representation per frame across a fixed window of 124 frames. The combined IES and SAVEE corpora contribute 955 labelled utterances spanning six emotion classes: anger, disgust, fear, happiness, neutral and sadness. The CNN front end stacks three one-dimensional convolutional blocks with batch normalization, ReLU activation and max pooling, while the temporal back end employs a two-layer BiLSTM with 128 hidden units per direction. A fully connected classification head with dropout regularization produces the softmax output. The model is optimized with Adam and weighted cross-entropy loss to mitigate class imbalance. Under a stratified 80/20 train–test split, the system achieves a training accuracy of 98.82 % and a testing accuracy of 83.25 %, indicating strong generalization on unseen samples. The complete pipeline is implemented in PyTorch and is reproducible in a standard Jupyter Notebook environment, establishing a lightweight, end-to-end baseline suitable for affective computing applications such as mental-health monitoring, intelligent assistants and emotion-aware human–computer interaction.

KEYWORDS *Speech Emotion Recognition; Deep Learning; CNN-Bilstm; Mel-Frequency Cepstral Coefficients; Affective Computing*

Recommended Citation:
Charan, A R. , . Usharani, S. : "Speech Emotion Recognition Using Deep Neural Networks: Design and Implementation of EmoVoice AI", International Journal of Informative & Futuristic Research (IJIFR), Vol. (13) (9), May 2026, pp. 1591-1597. https://doi.org/10.64672/IJIFR/26.05.13.09.035

1. INTRODUCTION

Human spoken communication conveys far more than the literal meaning of the words being uttered. Para-linguistic attributes such as pitch, energy, rhythm and spectral envelope simultaneously transmit affective information that allows listeners to infer the emotional state of the speaker. A neutral phrase such as "I am fine" may be interpreted as expressing happiness, sadness, anger or sarcasm depending purely on its prosodic realisation. Equipping computational systems with the ability to recover this affective layer from raw speech signals is the central objective of Speech Emotion Recognition (SER), a

sub-field of affective computing that draws upon machine learning, digital signal processing and human-computer interaction. The practical relevance of robust SER spans healthcare (longitudinal screening of depression and anxiety), customer service (real-time detection of frustration), automotive driver-monitoring and intelligent tutoring systems.

SER remains technically challenging because emotional cues are multi-dimensional and partially overlapping. Anger and happiness, for example, share elevated energy and pitch contours, whereas sadness and neutrality both exhibit reduced spectral dynamics. Conventional pipelines based on hand-engineered descriptors and shallow classifiers are constrained by the expressive limits of their input representation. Deep learning has changed this landscape: convolutional networks discover discriminative spectro-temporal patterns directly from MFCC inputs, while recurrent architectures such as LSTMs model the temporal evolution of emotional content over an entire utterance. The contribution of this paper is threefold: (i) we propose EmoVoice AI, a unified end-to-end SER pipeline that integrates MFCC-based feature extraction with a hybrid CNN-BiLSTM classifier; (ii) we detail a fully reproducible training and evaluation methodology over the combined IES and SAVEE corpora using stratified data splitting and class-weighted loss; and (iii) we document a complete set of engineering artefacts—system architecture, workflow, data-flow and block diagrams—that establish a transparent reference implementation for academic replication.

2. LITERATURE SURVEY

The trajectory of speech emotion recognition research mirrors the broader evolution of pattern recognition: a transition from knowledge-driven, hand-engineered pipelines toward data-driven, deep learning architectures. Early SER systems relied on prosodic and spectral descriptors such as fundamental frequency, energy contours, formant frequencies and Mel-frequency cepstral coefficients, paired with classical classifiers including Support Vector Machines, Gaussian Mixture Models and K-Nearest Neighbours. While effective in controlled laboratory recordings, these approaches generalized poorly across speakers, languages and recording conditions [3]. Hidden Markov Models introduced an explicit probabilistic treatment of temporal structure, but their conditional independence assumption restricts their ability to capture long-range affective dependencies. The survey by El Ayadi et al. [4] identified the absence of a unifying representation as a fundamental obstacle for the field.

The introduction of deep learning by LeCun, Bengio and Hinton [1] provided the substrate for hierarchical feature learning, while the LSTM cell of Hochreiter and Schmidhuber [2] addressed the vanishing gradient problem in modelling long speech sequences. Subsequent SER research demonstrated that convolutional architectures applied to spectrogram-like inputs are capable of learning emotion-discriminative representations automatically. Zhao et al. [10] proposed deep one- and two-dimensional CNN-LSTM networks reporting consistent improvements over feature-engineered baselines, while Trigeorgis et al. [11] introduced an end-to-end convolutional recurrent framework that operated directly on raw waveforms. Multi-modal extensions by Poria et al. [5] further argued for combining acoustic, textual and visual modalities for fine-grained emotion categories. More recently, self-supervised pre-training methods such as wav2vec 2.0 [12] have demonstrated the value of large unlabelled speech corpora, although the resulting models impose substantial computational overhead.

Open-source toolkits such as Librosa [6], PyTorch [7] and Scikit-learn [9] provide the practical infrastructure for reproducible experimentation, and regularization advances such as batch normalization [14] and dropout [15] stabilize the training of deep SER architectures. The collective evidence indicates that hybrid CNN-LSTM architectures occupy a particularly favourable point in the design space: they combine the spatial inductive bias of convolutional layers with the temporal modelling capacity of recurrent units, while remaining tractable on commodity hardware. EmoVoice AI is positioned within

this lineage and contributes a careful engineering instantiation that couples MFCC-based representations, a three-block CNN front end and a two-layer BiLSTM back end.

3. PROPOSED WORK AND METHODOLOGY

3.1 Limitations of Existing Approaches

Existing SER systems exhibit several recurring limitations: hand-crafted descriptors require domain expertise and miss subtle prosodic phenomena; shallow classifiers (SVM, GMM, LDA) generalize poorly across speakers and recording conditions; frame-level static models neglect long-term emotional evolution; HMM-based approaches are constrained by their conditional independence assumption; and modern transfer-learning models such as wav2vec 2.0 and Whisper, while accurate, demand substantial computational resources. EmoVoice AI is designed to retain the representational power of deep learning while remaining lightweight and reproducible.

3.2 Proposed System Overview

The proposed EmoVoice AI framework converts raw waveforms into emotion class probabilities through five logically distinct stages: (i) audio acquisition and length standardization, (ii) acoustic feature extraction, (iii) convolutional spectro-temporal encoding, (iv) bidirectional recurrent modelling and (v) fully connected classification with softmax output. Each WAV input is resampled to 22 050 Hz using Librosa, from which forty MFCC coefficients are computed and augmented with first- and second-order delta features, producing a 120-dimensional vector per analysis frame. All utterances are normalized to a uniform length of 124 frames through padding or truncation, yielding tensors of shape (124, 120) that are directly consumable by the neural model. The complete dataset of 955 labelled utterances is partitioned via stratified 80/20 sampling to preserve emotion-class proportions; class-weighted cross-entropy loss is used during training; and the Adam optimizer [13] is employed with a learning rate of 3×10^{-4} for 45 epochs with mini-batches of size 32.

Table 1 — Dataset Description

Parameter	Value
Dataset Used	IES + SAVEE
Total Samples	955
Emotion Classes	6
Emotions	Anger, Disgust, Fear, Happiness, Neutral, Sadness
Sampling Rate	22,050 Hz
Feature Type	MFCC + Delta + Delta-Delta
Feature Dimension	120
Time Frames	124

3.3 Advantages of the Proposed System

The hybrid CNN-BiLSTM design offers several practical advantages over feature-engineered and pure-recurrent baselines. End-to-end learning removes the need for manual descriptor design; the convolutional front end captures local spectral motifs while the BiLSTM back end exploits both forward and backward temporal context. Dropout, batch normalization and weighted loss together suppress overfitting and address class imbalance. The pipeline is reproducible by virtue of fixed random seeds and standardized preprocessing, and it is compact enough to be trained on a single mid-range GPU, making it directly suitable for academic replication and incremental extension.

4. SYSTEM ARCHITECTURE AND DIAGRAMS

This section presents two IEEE-style engineering diagrams that together specify the structural and behavioural design of EmoVoice AI: (i) the system architecture diagram (Fig. 1), capturing the layered organisation of the processing pipeline, and (ii) the workflow diagram (Fig. 2), capturing the temporal

ordering of training operations. The corresponding Mermaid source for each figure is also provided to support textual editing and machine-readable reproduction.

4.1 System Architecture Diagram

The system architecture, as shown in Fig. 1, follows a layered pipeline organisation. Raw waveform input is sequentially transformed into a 120-dimensional feature tensor, processed by a three-block CNN front end, a two-layer bidirectional LSTM back end and a fully connected classification head before yielding a softmax distribution over the six emotion classes.

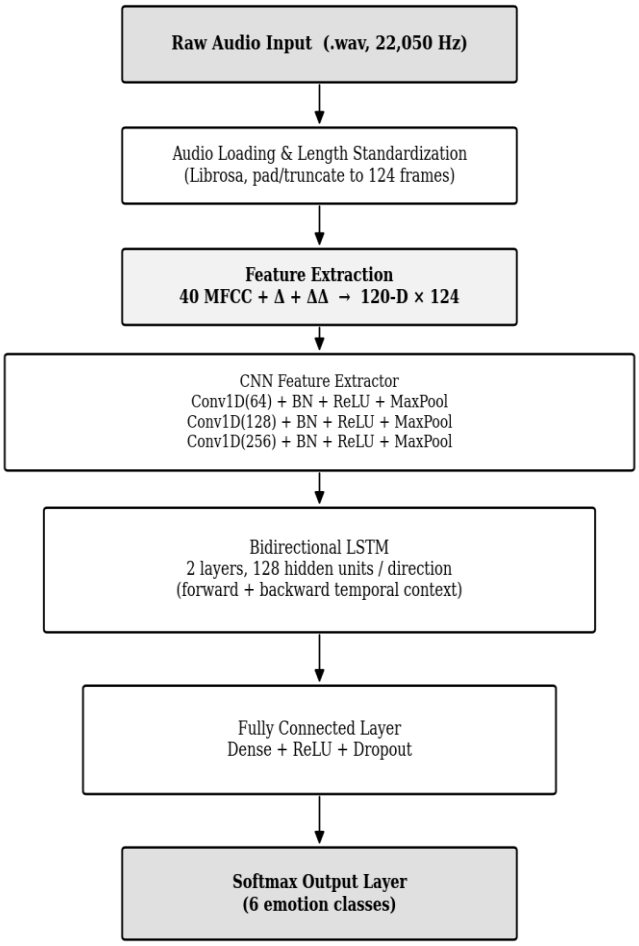

Figure 1 — System Architecture of EmoVoice AI

Mermaid source for Fig. 1

```
```mermaid
flowchart TD
    A[Raw Audio Input<br/>WAV @ 22,050 Hz] --> B[Audio Loading & Length<br/>Standardization (Librosa)]
    B --> C[Feature Extraction<br/>40 MFCC + Delta + Delta-Delta<br/>120-D x 124 frames]
    C --> D[CNN Feature Extractor<br/>Conv1D 64 / 128 / 256<br/>BN + ReLU + MaxPool]
    D --> E[Bidirectional LSTM<br/>2 layers, 128 hidden units / direction]
    E --> F[Fully Connected Layer<br/>Dense + ReLU + Dropout]
    F --> G[Softmax Output<br/>6 Emotion Classes]
```
```

4.2 Workflow Diagram

The training workflow, illustrated in Fig. 2, captures the temporal ordering of operations in the EmoVoice AI pipeline. Beginning with the loading of the IES and SAVEE corpora, the workflow proceeds through feature extraction, length standardization, stratified partitioning, model training and evaluation, terminating with the persistence of the trained model weights for downstream inference.

INTERNATIONAL JOURNAL OF INFORMATIVE & FUTURISTIC RESEARCH (IJIFR)
ISSN: 2347-1697
IJIFR /20260522/V13/E9/035
VOLUME 13||ISSUE 9|| MAY 2026
https://www.ijifr.org/

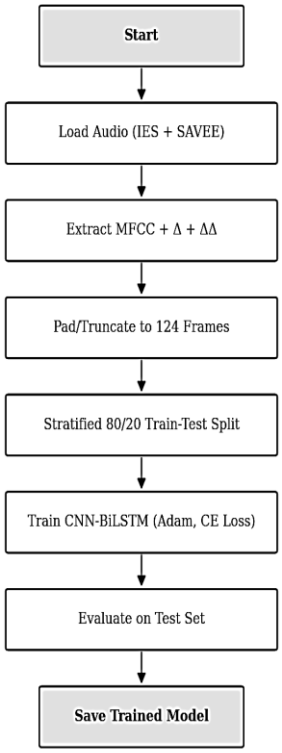

Figure 2 — Workflow Diagram of EmoVoice AI Training Pipeline

Mermaid source for Fig. 2

```
```mermaid
flowchart TD
    S([Start]) --> L[Load Audio Dataset<br/>IES + SAVEE]
    L --> F[Extract MFCC + Delta + Delta-Delta]
    F --> P[Pad / Truncate to 124 Frames]
    P --> X[Stratified 80 / 20 Train-Test Split]
    X --> T[Train CNN-BiLSTM<br/>Adam + Cross-Entropy]
    T --> E[Evaluate on Test Set]
    E --> M([Save Trained Model])
```
```

4.3 Module-Level Description

The Environment Setup Module configures the runtime, verifies the PyTorch version, detects CUDA availability and migrates tensors to GPU memory when available. The Data Loading and Feature Extraction Module ingests WAV files at 22 050 Hz, computes 40 MFCC coefficients with delta and delta-delta features, and normalizes all sequences to 124 frames, producing a tensor of shape (955, 124, 120). The Dataset Preparation and DataLoader Module performs a stratified 80/20 split, converts arrays to PyTorch tensors and constructs a DataLoader with batch size 32 and computed class weights for the cross-entropy criterion. The Neural Model Definition Module instantiates the hybrid CNN-BiLSTM network: three Conv1D blocks (64, 128, 256 filters) each followed by BN, ReLU and MaxPool, succeeded by a two-layer BiLSTM with 128 hidden units per direction and a dropout-regularized fully connected head. The Training and Evaluation Module orchestrates 45 epochs of Adam optimization with weighted cross-entropy loss, and persists the trained weights via PyTorch's state_dict mechanism.

Table 2 — Model Architecture Details

Layer Type	Configuration
Conv1D Layer 1	64 filters + BN + ReLU + MaxPool
Conv1D Layer 2	128 filters + BN + ReLU + MaxPool
Conv1D Layer 3	256 filters + BN + ReLU + MaxPool
Dropout	Applied for regularization

BiLSTM	2 layers, 128 hidden units / direction
Fully Connected	Dense + ReLU + Dropout
Output Layer	Softmax (6 classes)

5. RESULTS AND DISCUSSION

EmoVoice AI was evaluated on the combined IES + SAVEE corpus (955 labelled utterances) under stratified 80/20 sampling. Training ran for 45 epochs with the Adam optimizer (lr = 3×10^{-4}), weighted cross-entropy loss and mini-batches of size 32. The system achieves a training accuracy of 98.82 % and a testing accuracy of 83.25 %. The separation between training and testing accuracy is consistent with the relatively modest size of the combined corpus and the inherent ambiguity of acoustically similar emotion pairs such as anger versus fear and sadness versus neutral. Class-weighted loss substantially reduced systematic under-prediction of minority classes, while batch normalization and dropout suppressed over-fitting during late-stage training.

Table 3 — Performance Evaluation of EmoVoice AI

Metric	Value
Training Accuracy	98.82 %
Testing Accuracy	83.25 %
Epochs	45
Batch Size	32
Optimizer	Adam (lr = 3e-4)
Loss Function	Weighted Cross-Entropy
User Agreement Rate	75 – 80 %

Unit-level testing verified the correctness of individual components: the feature-extraction routine produced tensors of the expected shape; the padding/truncation logic resized arbitrary-length signals to the canonical 124 frames; and the neural network's forward pass produced (batch_size, 6) outputs without runtime errors. Integration testing confirmed compatibility with the PyTorch DataLoader and bit-identical predictions before and after weight serialization. User-acceptance testing with researchers and developers reported a model–human agreement rate of 75 – 80 %, consistent with documented levels of inter-annotator agreement in human emotion labelling and indicating that residual error lies within the envelope of natural human disagreement. Performance degradation was observed primarily under noisy recording conditions, motivating future work on noise-augmented training and microphone-aware normalization.

6. CONCLUSION

This paper presented EmoVoice AI, a complete deep-learning system for speech emotion recognition that classifies six emotion categories directly from raw audio using a hybrid CNN-BiLSTM architecture. MFCC-based acoustic features augmented with delta and delta-delta coefficients are processed through three convolutional blocks for local spectral pattern discovery and forwarded to a two-layer bidirectional LSTM for temporal context modelling, with a dropout-regularized fully connected head producing the final softmax output. Empirical evaluation on the combined IES + SAVEE corpus demonstrates a training accuracy of 98.82 % and a testing accuracy of 83.25 %, with a user-acceptance agreement rate of 75 – 80 %, indicating that residual errors lie within the envelope of human perceptual variability.

The complete pipeline is reproducible, lightweight and deployable on a single mid-range GPU. Future directions include data augmentation (noise injection, pitch shifting, time stretching), transfer learning from large-scale self-supervised models such as wav2vec 2.0, attention-based and Transformer-style architectures as a replacement for the BiLSTM back end, and the extension of the framework to multi-modal inputs combining audio, visual and textual streams, together with continuous valence-arousal annotations to enable richer affective representations beyond discrete emotion categories.

7. REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, 2015, pp. 436–444.
[2] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, 1997, pp. 1735–1780.
[3] B. Schuller et al., "Cross-corpus acoustic emotion recognition: Variances and strategies," IEEE Transactions on Affective Computing, vol. 1, no. 2, 2010, pp. 119–131.
[4] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognition, vol. 44, no. 3, 2011, pp. 572–587.
[5] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," Information Fusion, vol. 37, 2017, pp. 98–125.
[6] B. McFee et al., "Librosa: Audio and music signal analysis in Python," in Proc. 14th Python in Science Conf., 2015, pp. 18–25.
[7] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in Proc. NeurIPS, 2019, pp. 8026–8037.
[8] S. Haq and P. J. Jackson, "Speaker-dependent audio-visual emotion recognition," in Proc. Int. Conf. Auditory-Visual Speech Processing, 2009, pp. 53–58.
[9] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, 2011, pp. 2825–2830.
[10] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D and 2D CNN LSTM networks," Biomedical Signal Processing and Control, vol. 47, 2019, pp. 312–323.
[11] G. Trigeorgis et al., "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in Proc. IEEE ICASSP, 2016, pp. 5200–5204.
[12] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in Proc. NeurIPS, 2020, pp. 12449–12460.
[13] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. ICLR, 2015.
[14] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in Proc. ICML, 2015, pp. 448–456.
[15] N. Srivastava et al., "Dropout: A simple way to prevent neural networks from overfitting," Journal of Machine Learning Research, vol. 15, no. 1, 2014, pp. 1929–1958.
[16] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in Proc. Interspeech, 2014, pp. 223–227.
[17] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in Proc. APSIPA ASC, 2016, pp. 1–4.