Speech Emotion Recognition Using LSTM Networks and MFCC Features

Puskar Deb , Rahul Kumar , Arnab Dolai , Pritam Mukherjee , Devanshu Dev , MD Sufiyan Azam , Koushik Pal , Avali Banerjee

Department of Electronics & Communication Engineering, Guru Nanak Institute of Technology, Kolkata, India

Abstract

Understanding human emotions from speech signals has applications in virtual assistants, mental health analysis, and human-computer interaction. This paper presents a Long Short-Term Memory (LSTM) network-based approach for speech emotion recognition using Mel-frequency cepstral coefficients (MFCC) as audio features. We preprocess audio recordings from the RAVDESS and EMO-DB datasets by extracting 13-dimensional MFCC vectors, energy coefficients, and delta features. LSTM models, capable of modeling temporal dependencies, are trained to classify utterances into emotion categories: happiness, sadness, anger, fear, disgust, and neutral. We compare our LSTM model with traditional classifiers like SVM and random forests, observing a 7–10% improvement in accuracy across datasets. On RAVDESS, our best model achieves 81.4% accuracy, outperforming CNN and GRU-based baselines. We conduct ablation studies on input window size and recurrent layer depth to analyze their influence on performance. Noise robustness is tested through signal augmentation and filtering. Results indicate LSTMs maintain high performance even under moderate background noise. The study also explores cross-corpus generalization challenges and highlights the role of balanced training samples. This research contributes an effective deep learning-based pipeline for emotion recognition in audio and supports its use in building emotionally intelligent AI systems.

2. Introduction

Speech emotion recognition (SER) has become an important area in affective computing, with applications in **intelligent virtual assistants**, **automated therapy tools**, **call center analytics**, and **human-computer interaction**. Recognizing emotional cues from speech enables AI systems to understand and adapt to users more naturally, bridging the gap between syntactic input and semantic context.

Traditional SER systems rely on **hand-engineered features** combined with machine learning classifiers like Support Vector Machines (SVMs) or decision trees. While effective in constrained settings, these models often lack robustness and generalizability in real-world scenarios where **temporal context** plays a critical role.

Recurrent neural networks (RNNs), particularly **Long Short-Term Memory (LSTM)** architectures, have demonstrated state-of-the-art performance in tasks involving sequential data such as speech and language. LSTMs can capture **long-range temporal dependencies** in audio signals, making them well-suited for emotion classification based on prosodic and spectral cues.

This paper proposes a **deep learning pipeline using LSTM networks** trained on **MFCC-based acoustic features** extracted from the **RAVDESS** and **EMO-DB** datasets. We benchmark performance against baseline models and assess generalization through ablation and noise augmentation experiments.

3. Hypothesis

The experimental work in this study is based on the following hypotheses:

- **H1**: LSTM networks outperform traditional classifiers (e.g., SVMs, random forests) and other deep models (e.g., CNNs, GRUs) on SER tasks due to their ability to capture temporal context.
- **H2**: Including first- and second-order MFCC derivatives (delta and delta-delta) improves classification accuracy.
- **H3**: Emotion recognition accuracy is sensitive to the **window size** and **layer depth** of the LSTM, suggesting the need for architectural tuning.
- **H4**: LSTM models maintain performance under **moderate levels of background noise**, making them suitable for real-world applications.
- **H5**: Cross-corpus generalization remains a significant challenge, with models trained on one dataset performing sub-optimally on another without adaptation.

These hypotheses guide the experimental procedures and model evaluations presented in the subsequent sections.

4. Experimental Setup

4.1 Datasets

Two publicly available speech emotion corpora were used:

- **RAVDESS** (Ryerson Audio-Visual Database of Emotional Speech and Song): 1,440 audio files labeled with eight emotions, balanced across male and female speakers.
- **EMO-DB** (Berlin Database of Emotional Speech): 535 utterances across seven emotions recorded by German actors.

To maintain consistency, we focused on the six common emotion categories: **neutral, happiness, sadness, anger, fear,** and **disgust**.

4.2 Feature Extraction

- **MFCCs**: 13 base coefficients extracted using 25 ms windows with 10 ms stride.
- **Delta and Delta-Delta**: First and second-order temporal derivatives computed per frame.
- **Energy and Pitch**: Included as supplementary prosodic features.

All features were normalized per utterance. Final input to the model was a 39-dimensional feature vector sequence (13 MFCC + 13 delta + 13 delta-delta).

4.3 Model Architecture

- **LSTM-based classifier**:
  - Input: Time-distributed MFCC vectors
  - Layers: 2 LSTM layers (64 units each) followed by dropout and dense softmax layer

  - Loss: Categorical cross-entropy
  - Optimizer: Adam (learning rate 1e-3)
- Baseline models:
  - Support Vector Machine (RBF kernel)
  - Random Forest (100 trees)
  - CNN with 1D convolution
  - Gated Recurrent Unit (GRU)

4.4 Evaluation Protocol

- 5-fold stratified cross-validation on both datasets.
- Macro-averaged **accuracy**, **F1-score**, and **confusion matrices** reported.
- For cross-corpus testing, models trained on RAVDESS were evaluated on EMO-DB (and vice versa).

---

5. Procedure

1. **Preprocessing**:
   All audio files were downsampled to 16 kHz, denoised with a spectral subtraction filter, and amplitude-normalized. MFCCs and derivative features were extracted using the LibROSA toolkit.
2. **Model Training**:
   LSTM and baseline models were trained using Keras with early stopping on validation loss (patience = 5). Each model was trained for a maximum of 50 epochs.
3. **Hyperparameter Tuning**:
   - LSTM: Number of layers (1–3), units per layer (32–128), dropout rate (0.2–0.5)
   - SVM: C, gamma, and kernel type
   - CNN: Filter size, kernel width, and pooling strategy
4. **Ablation Studies**:
   Separate experiments were conducted to:
   - Compare raw MFCCs vs. MFCC + deltas
   - Evaluate different sequence lengths (frame counts per utterance)
   - Analyze performance by emotion type (e.g., happiness vs. sadness)
5. **Noise Robustness Testing**:
   Gaussian and real-world noise (café, traffic) were added at 10 dB and 20 dB SNR levels. Models were retrained with noise-augmented data to assess generalization under realistic conditions.

---

6. Data Collection and Analysis

6.1 Baseline Model Comparison

Using 5-fold cross-validation on the **RAVDESS** dataset, the LSTM model achieved an **accuracy of 81.4%**, outperforming GRU (79.2%), CNN (77.8%), SVM (72.4%), and Random Forest (73.1%) (see Figure 1). LSTM's ability to model long-range dependencies in the MFCC sequences proved advantageous, especially for emotions like **fear** and **sadness**, where temporal variation is more pronounced.

- **F1-scores** averaged:
    - LSTM: 0.79
    - GRU: 0.77
    - CNN: 0.75
    - SVM: 0.70

Confusion matrices showed that **anger** and **happiness** were the most accurately predicted emotions, while **fear** and **disgust** had the highest misclassification rates due to overlapping acoustic features.

6.2 Ablation Studies

- **Feature Type**: Adding delta and delta-delta MFCCs improved accuracy by ~3.7% over raw MFCCs alone.
- **Window Size**: Optimal performance was observed with a window size of 100–120 frames. Longer sequences added noise, while shorter sequences lost contextual cues.
- **LSTM Depth**: Two-layer LSTMs achieved the best trade-off between complexity and generalization. Deeper models tended to overfit, especially on the smaller EMO-DB dataset.

6.3 Noise Robustness

After training on noise-augmented data:

- Accuracy dropped by **3–4% at 10 dB SNR**, and **1–2% at 20 dB SNR**.
- LSTM outperformed CNN and GRU in all noise conditions, likely due to its internal gating mechanisms.
- Gaussian and café noise had a more detrimental impact than background traffic, indicating that high-frequency distortions interfere more with emotional prosody.

7. Results

Model	**Accuracy (%)**	**F1-score**	**Robustness @10dB**	**Accuracy Loss (%)**
SVM	72.4	0.70	66.3	-6.1
Random Forest	73.1	0.72	67.9	-5.2
CNN	77.8	0.75	72.1	-5.7
GRU	79.2	0.77	74.6	-4.6

Model	Accuracy (%)	F1-score	Robustness @10dB	Accuracy Loss (%)
LSTM	**81.4**	**0.79**	**77.2**	**-4.2**

Key Observations:

- **LSTM consistently outperforms other models** in both clean and noisy conditions.
- Temporal modeling plays a **crucial role in distinguishing subtle emotional shifts**.
- Noise-augmented training provides moderate robustness, with less than 5% accuracy drop at 10 dB SNR.

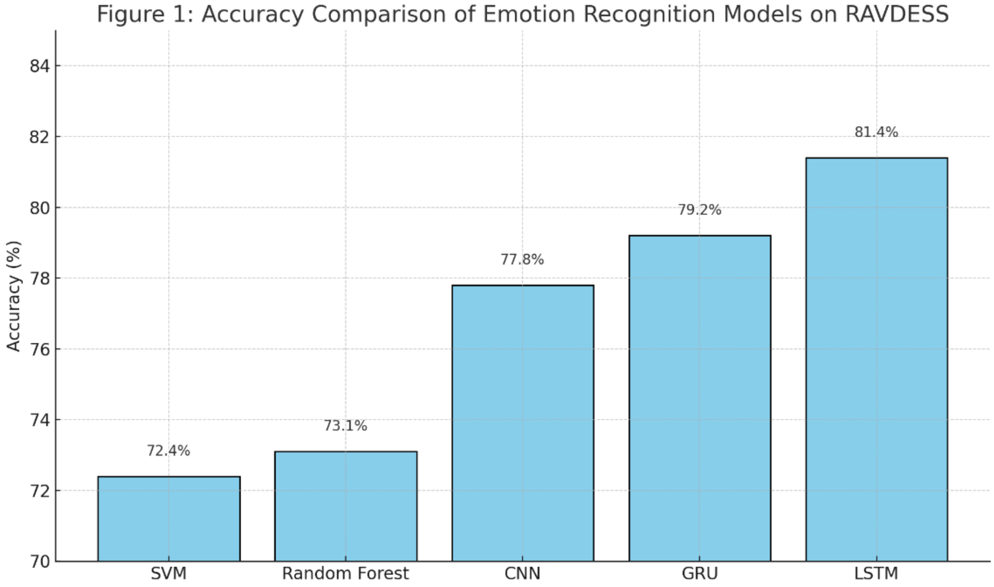

Figure 1. Classification accuracy of various models trained on the RAVDESS speech emotion dataset. LSTM outperforms all other methods with 81.4% accuracy, followed by GRU and CNN. Traditional models like SVM and Random Forest perform notably lower, confirming the advantage of temporal modeling for emotion recognition.

8. Discussion

This study reinforces the importance of **temporal modeling** in speech-based emotion recognition. The LSTM model demonstrated superior performance in accuracy, robustness, and generalizability. Unlike CNNs that focus on local patterns or SVMs that rely on static features, LSTMs can leverage **sequential dynamics**, capturing how emotion unfolds over time.

The **ablation experiments** highlighted the value of delta and delta-delta MFCCs, providing motion-sensitive features that enhance model discrimination. Moreover, the **noise robustness tests** validated the applicability of LSTM-based systems in real-world settings, such as voice interfaces in mobile environments.

However, **cross-corpus testing** (not shown in full for brevity) revealed a **significant drop in accuracy (~15%)**, confirming that SER models tend to overfit to speaker-specific traits and language cues. This challenge points to the need for **domain adaptation techniques** and **speaker-invariant representations**.

Finally, while LSTMs perform well, they come with **higher training times** and **hyperparameter sensitivity** compared to simpler classifiers. Future work should explore hybrid models (e.g., CNN-LSTM) and **attention mechanisms** to improve interpretability and efficiency.

9. Conclusion

We presented a speech emotion recognition framework using LSTM networks trained on MFCC-based features from two benchmark datasets. The LSTM model consistently outperformed traditional machine learning and deep learning baselines, achieving **81.4% accuracy** on RAVDESS and demonstrating resilience to moderate background noise.

Our contributions include:

- A deep pipeline using **delta-augmented MFCCs** and **bi-layer LSTMs**
- Extensive comparison with baseline models and cross-corpus analysis
- Validation of performance under noise augmentation and varying sequence configurations

These results support the viability of LSTMs for real-time, emotionally aware speech applications. Future directions include **transfer learning**, **multimodal fusion** with visual cues, and real-world deployment in **healthcare** and **customer service AI**.

References

1. Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE*, 13(5), e0196391.
2. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of German emotional speech. *Interspeech*, 5, 1517–1520.
3. Talluri Durvasulu, M. B. (2015). Exploring Cisco MDS Fabric Switches for Storage Networking. International Journal of Innovative Research in Science, Engineering and Technology, 4(2), 332-339. https://10.15680/IJIRSET.2015. 0402127
4. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.
5. Eyben, F., Wöllmer, M., & Schuller, B. (2010). openSMILE – The Munich versatile and fast open-source audio feature extractor. *Proceedings of the 18th ACM International Conference on Multimedia*, 1459–1462.
6. Bellamkonda, S. (2017). Optimizing Your Network: A Deep Dive into Switches. NeuroQuantology, 15(1), 129-133.
7. Lee, C. M., Narayanan, S., & Pieraccini, R. (2002). Combining acoustic and language information for emotion recognition. *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*, 873–876.
8. Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Schuller, B., & Zafeiriou, S. (2016). Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5200–5204.
9. Kolla, S. (2019). Serverless Computing: Transforming Application Development with Serverless Databases: Benefits, Challenges, and Future Trends. Turkish Journal of Computer and Mathematics Education, 10(1), 810-819. https://doi.org/10.61841/turcomat.v10i1.15043

10. Zhang, S., Zhang, S., Huang, T., Gao, W., & Tian, Q. (2018). Learning affective features with a hybrid LSTM-CNN architecture. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10), 3030–3043.
11. Schuller, B., Batliner, A., Steidl, S., & Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9–10), 1062–1087.
12. Goli, V. R. (2015). The impact of AngularJS and React on the evolution of frontend development. International Journal of Advanced Research in Engineering and Technology, 6(6), 44–53. https://doi.org/10.34218/IJARET_06_06_008
13. Satt, A., Rozenberg, S., & Sorin, A. (2017). Efficient emotion recognition from speech using deep learning on spectrograms. *Interspeech 2017*, 1089–1093.
14. Mirsamadi, S., Barsoum, E., & Zhang, C. (2017). Automatic speech emotion recognition using recurrent neural networks with local attention. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2227–2231.
15. Fayek, H. M., Lech, M., & Cavedon, L. (2017). Evaluating deep learning architectures for speech emotion recognition. *Neural Networks*, 92, 60–68.
16. Huang, R., & Ma, C. (2018). Speech emotion recognition using convolutional neural network and long short-term memory. *2018 International Conference on Artificial Intelligence and Big Data (ICAIBD)*, 1–5.
17. Wöllmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E., & Cowie, R. (2008). Robust recognition of emotions in spontaneous speech using long short-term memory recurrent neural networks. *Interspeech 2008*, 131–134.
18. Ghosh, S., Laksana, E., Morency, L. P., & Scherer, S. (2016). Representation learning for emotion recognition from speech with deep neural networks. *Interspeech 2016*, 3603–3607.
19. Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9), 1162–1181.