



A Deep Learning Framework for Real-Time Speech Emotion Recognition

Dr. N S Patil¹, Deeksha S Gowda², Harsha R Nadigar³, Hruthwik C⁴, Keerthana K S⁵

¹Associate Professor, Department of Information Science and Engineering, Bapuji Institute of Engineering, Karnataka, India.

^{2,3,4,5}Department of Information Science and Engineering, Bapuji Institute of Engineering, Karnataka, India.

To Cite this Article: Dr. N S Patil¹, Deeksha S Gowda², Harsha R Nadigar³, Hruthwik C⁴, Keerthana K S⁵, "A Deep Learning Framework for Real-Time Speech Emotion Recognition", Indian Journal of Computer Science and Technology, Volume 05, Issue 03 (September-December 2026), PP: 537-542.



Copyright: ©2026 This is an open access journal, and articles are distributed under the terms of the [Creative Commons Attribution License](#); Which Permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract: Speech Emotion Recognition (SER) is the task of estimating a speaker's emotion from their voice. This study compares three ways of representing speech using Mel-Frequency Cepstral Coefficients (MFCCs): 13, 26, and 39 coefficients. The features are supplied to a Stacked Gated Recurrent Unit (GRU) model, which learns patterns that change over the course of an utterance. The experiments use 4,240 audio samples from the RAVDESS and Toronto Emotional Speech Set (TESS) datasets, grouped into seven emotion classes. Each MFCC sequence is padded or shortened to 200 frames, and feature scaling is fitted using the training data only. The three models use the same architecture and training settings so their results can be compared on a consistent basis. Evaluation in the project includes classification metrics, one-vs-rest ROC-AUC, probability calibration curves, and t-SNE plots. The Phase-II report identifies MFCC-39 as the most promising of the three variants and MFCC-13 as a lighter baseline. Since the report's summary table labels its accuracy and F1 values as expected performance, those estimates should not be presented as confirmed test results unless supported by the final experimental outputs.

Key Word: Speech Emotion Recognition, MFCC, Stacked GRU, Deep Learning, RAVDESS, TESS, ROC-AUC, t-SNE.

I. INTRODUCTION

A person's voice carries information beyond the words being spoken. Pitch, loudness, rhythm, and changes in the sound spectrum can provide clues about emotions such as happiness, sadness, anger, and fear. Speech Emotion Recognition (SER) uses these acoustic patterns to estimate the emotional category expressed in a recording. It may be useful in applications such as conversational assistants, customer-service analysis, and assistive technologies, although reliable use outside controlled datasets requires careful testing. Earlier SER systems often combined manually selected acoustic features—such as pitch, energy, zero-crossing rate, and MFCCs—with classifiers including Support Vector Machines, Gaussian Mixture Models, or Hidden Markov Models. These methods can work well in controlled settings, but their performance may change when speakers or recording conditions differ from those used during training. Deep-learning methods offer another approach: recurrent models can learn patterns across successive audio frames rather than treating each frame independently. In this project, a Stacked GRU model is used to study the effect of changing the number of MFCC coefficients. The main question addressed in this study is whether MFCC dimensionality affects the performance of a GRU-based SER model when the remaining experimental settings are held constant. We compare 13, 26, and 39 coefficients and examine more than accuracy alone. Precision, Recall, F1-score, ROC-AUC, calibration curves, and t-SNE visualisations are included to provide a broader view of classification behaviour.

II. SCOPE OF THE PROJECT

The project focuses on developing and evaluating a deep-learning framework for recognising emotions from labelled speech recordings. It combines two publicly available speech emotion datasets, RAVDESS and TESS, and uses MFCC features as the input representation for a Stacked GRU classifier. The system compares three feature configurations containing 13, 26, and 39 MFCC coefficients. Each audio sequence is adjusted to 200 frames, and the resulting features are standardised before model training. Three independent models are trained under the same experimental settings so that their performance can be compared fairly. The project covers dataset preparation, feature extraction, preprocessing, model development, training, and performance evaluation. It also examines confusion matrices, ROC curves, probability calibration, and t-SNE visualisations.

III. EXISTING SYSTEM AND PROPOSED SYSTEM

Existing System: Existing speech emotion recognition systems commonly use manually selected acoustic features followed by conventional machine-learning classifiers. Features such as pitch, energy, zero-crossing rate, and MFCCs are extracted from audio signals and supplied to algorithms such as Support Vector Machines, Gaussian Mixture Models, and Hidden Markov Models. These systems can be computationally efficient and useful for controlled datasets. However, selecting appropriate features often requires domain knowledge, and performance may be affected by differences in speakers, accents, background noise, and recording conditions. Deep-learning methods, including Convolutional Neural Networks, Long Short-

Term Memory networks, and Gated Recurrent Units, have provided additional ways to learn patterns from speech signals. CNNs can identify spectral patterns, while recurrent networks can model changes in acoustic features over time. Despite these developments, the choice of MFCC dimensionality remains an important consideration. Using fewer coefficients reduces the size of the input representation, whereas using more coefficients may preserve additional acoustic information. A controlled comparison is therefore useful for understanding how this choice affects a GRU-based SER model.

Proposed System: The proposed system uses a Stacked GRU architecture to classify emotions from sequential MFCC features. It combines 1,440 audio samples from RAVDESS and 2,800 samples from TESS, resulting in a dataset of 4,240 samples. The dataset is organised into seven emotion categories: neutral, happy, sad, angry, fear, disgust, and surprise. MFCC features are extracted using three configurations: 13, 26, and 39 coefficients. Each sequence is padded or truncated to 200 frames to maintain a consistent input length. Three separate GRU models are developed using the same architecture and training settings. Each model contains recurrent layers with 256, 128, and 64 units, followed by Batch Normalisation, Dropout, a Dense layer, and a Softmax output layer. The dataset is divided into training, validation, and testing subsets using a stratified 60:20:20 split. The models are trained using the Adam optimiser and evaluated using several classification and visualisation techniques. By keeping the model architecture and training procedure consistent, the study aims to compare the effect of MFCC dimensionality without introducing unnecessary differences between the experimental configurations.

IV. SYSTEM DESIGN

System design: The proposed framework follows a sequence of processing stages, beginning with audio dataset preparation and ending with model evaluation. Each stage contributes to the overall process of identifying emotions from speech recordings.

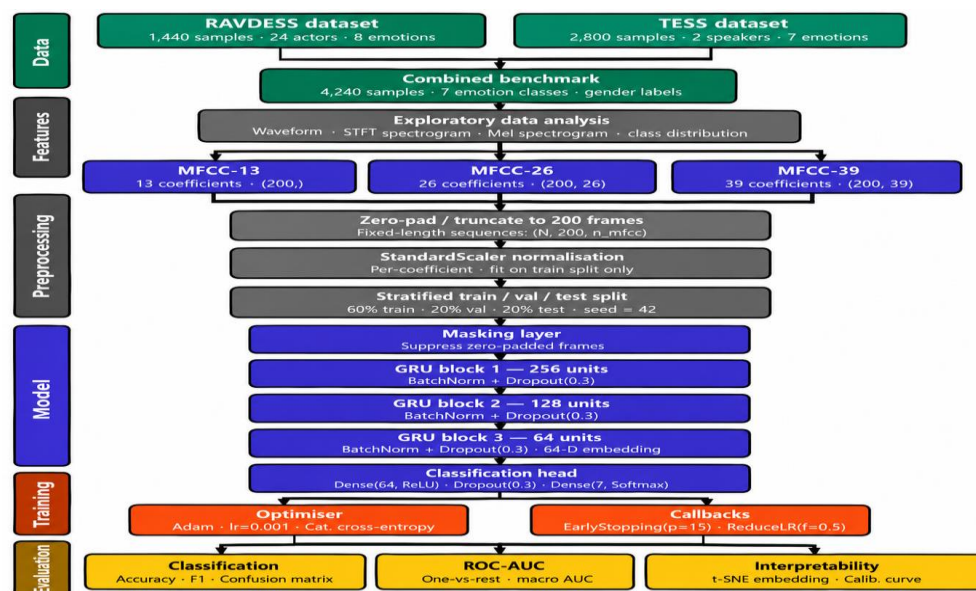


Figure no 1: System Design[

1. **Dataset Preparation:** The project uses two speech emotion datasets: RAVDESS and TESS. RAVDESS contains 1,440 audio samples recorded by 24 actors. TESS contains 2,800 audio samples from two female speakers. The two datasets provide labelled speech recordings representing different emotional states. To create a consistent classification task, the calm category in RAVDESS is merged with the neutral category. The combined dataset contains seven emotion classes: neutral, happy, sad, angry, fear, disgust, and surprise. The resulting dataset contains 4,240 audio samples. The combination of the two datasets provides a common basis for comparing the three MFCC configurations.
2. **MFCC Feature Extraction:** Mel-Frequency Cepstral Coefficients are widely used in speech processing because they represent important characteristics of the short-term frequency spectrum of an audio signal. In this project, the Librosa library is used to load audio files and extract MFCC features. Three configurations are considered: MFCC-13, MFCC-26, and MFCC-39. The purpose of comparing these configurations is to investigate whether increasing the number of coefficients provides additional information that helps the model distinguish between emotions.
3. **Feature Standardisation:** After feature extraction, the MFCC coefficients are standardised before being supplied to the models. Standardisation helps place features on comparable scales and can support more stable model training. The mean and standard deviation are calculated using only the training data. The same transformation is then applied to the validation and test datasets.
4. **Stacked GRU Model Architecture:** The proposed model uses a Stacked Gated Recurrent Unit architecture to learn temporal patterns from MFCC sequences. The first GRU layer contains 256 units and processes the input sequence. The second GRU layer contains 128 units, followed by a third GRU layer containing 64 units. These layers progressively transform the acoustic sequence into a compact representation that can be used for emotion classification. A Masking layer is included to handle padded sequences. Batch Normalisation is used between recurrent stages, while Dropout with a rate of 0.3 is applied to help reduce overfitting.

5. **Model Training:** The dataset is divided into training, validation, and testing subsets using a stratified ratio of 60:20:20. The training set is used to train the Stacked GRU model, while the validation set monitors performance during training and the testing set evaluates the model on unseen data. The models use the Adam optimiser with a learning rate of 0.001, categorical cross-entropy loss, and a batch size of 32. Training runs for a maximum of 280 epochs. EarlyStopping monitors validation loss and stops training if it does not improve for 15 consecutive epochs. ReduceLROnPlateau reduces the learning rate when validation loss stops improving, helping the model achieve better convergence.
6. **Performance Evaluation:** The trained models are evaluated using several complementary measures. **Accuracy** measures the proportion of correctly classified samples. Precision indicates how many samples predicted as a particular emotion belong to that class, while Recall measures how many actual samples of that emotion are correctly identified. The F1-score combines Precision and Recall into a single measure. Confusion matrices are used to examine the classification errors between emotion categories. They help identify which emotions the model distinguishes successfully and which categories are more frequently confused. ROC curves and ROC-AUC are calculated using a one-vs-rest approach to examine how well the models distinguish individual emotion classes from the remaining classes.

Finally, t-SNE is applied to the learned feature embeddings to visualise how the model represents different emotion categories. These visualisations provide an additional perspective on the separation of the learned representations.

V. RESULT

The proposed Speech Emotion Recognition (SER) system is evaluated using three Stacked GRU models with 13, 26, and 39 MFCC coefficients. The models are trained and tested using the combined RAVDESS and TESS datasets, containing 4,240 speech samples classified into seven emotion categories. Performance is assessed using accuracy, precision, recall, F1-score, and ROC-AUC. Based on the expected performance described in the project report, the models are estimated to achieve approximately 80%, 83%, and 90% accuracy for 13, 26, and 39 MFCC coefficients, respectively. The corresponding expected macro F1-scores are 0.80, 0.84, and 0.87. These estimates suggest that increasing the number of MFCC coefficients may improve the model's ability to capture important speech characteristics for emotion classification. The evaluation also considers precision and recall to examine classification performance across different emotions. ROC-AUC can provide additional insight into the model's ability to distinguish between emotion classes, while confusion matrices can help identify emotions that are frequently misclassified.

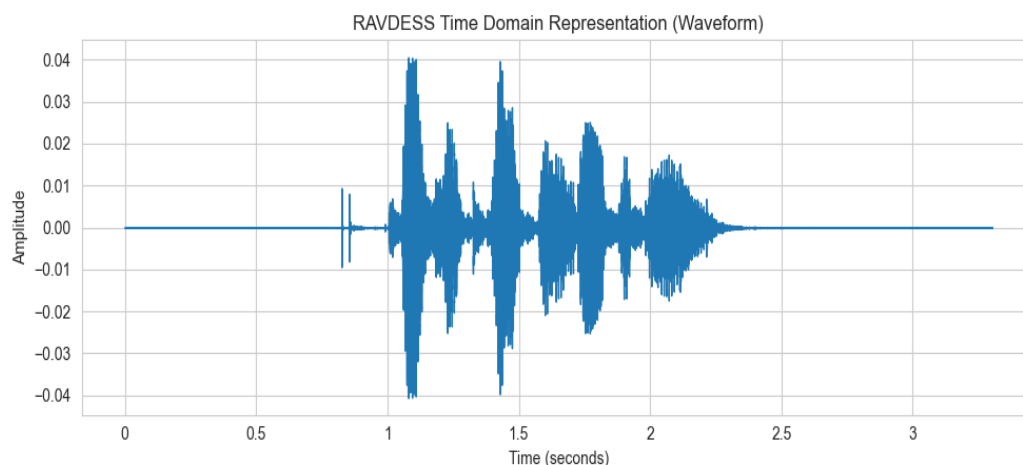


Figure no 1: RAVDESS waveform — time domain (neutral)

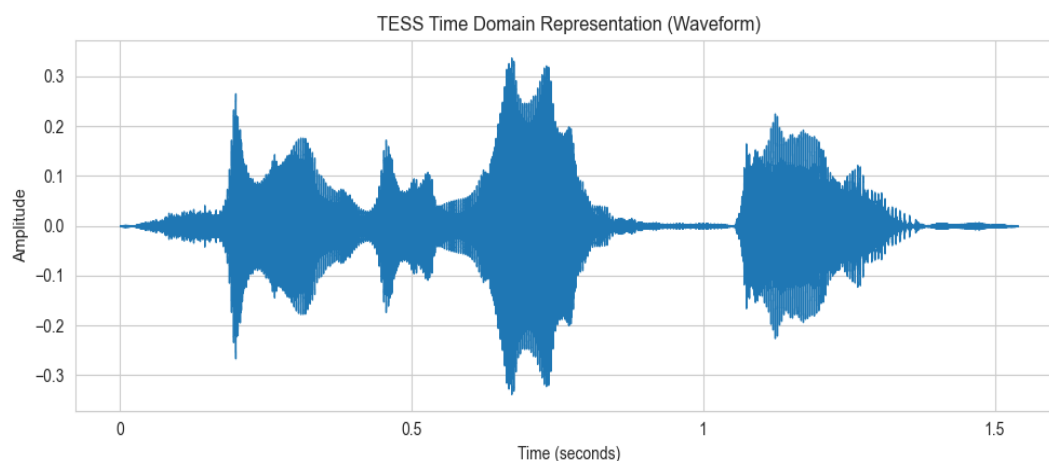


Figure no 2: TESS waveform — time domain (angry)

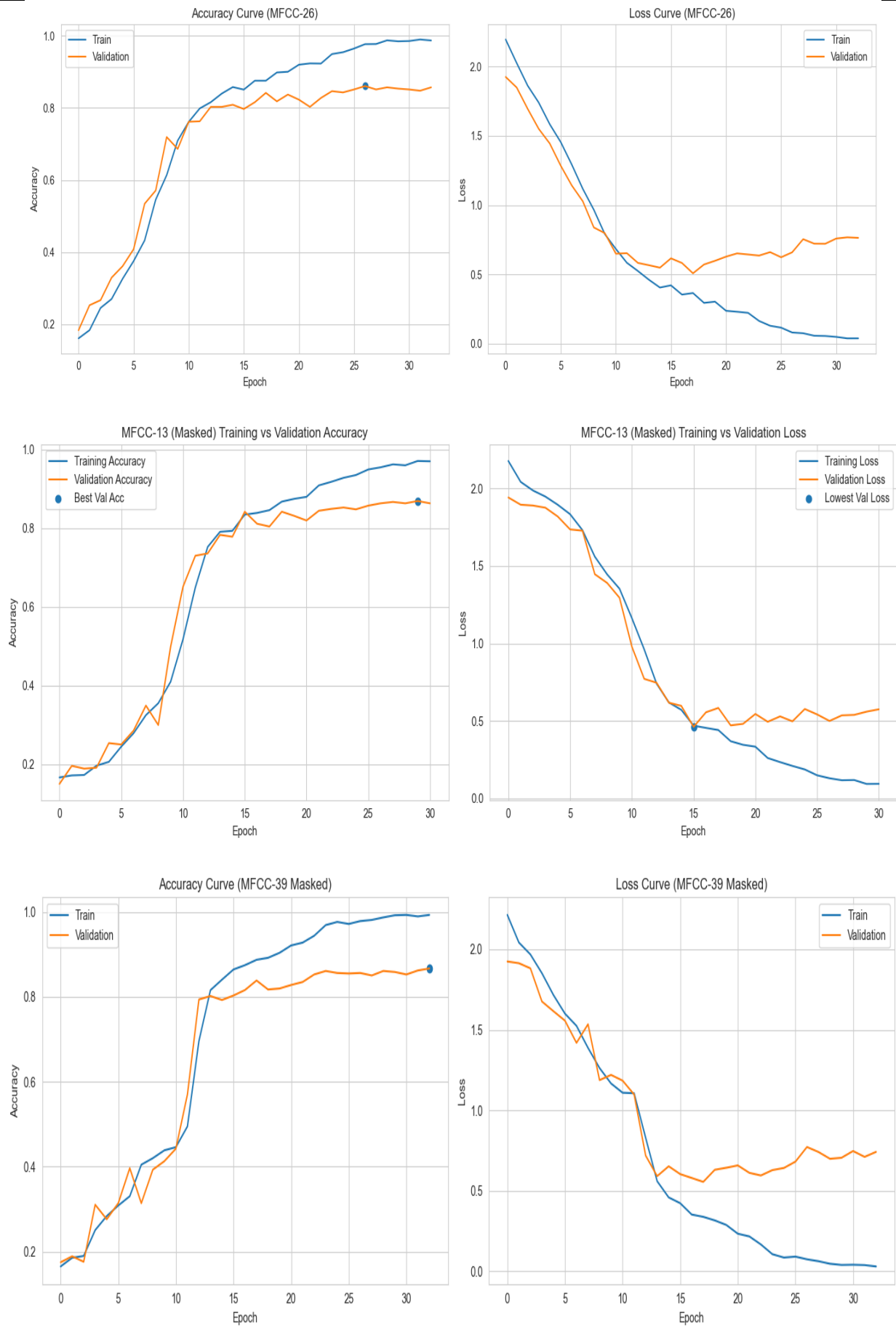


Figure no 3: MFCC-13 — best val. accuracy 0.8691 at epoch 30
 MFCC-26 — best val. accuracy 0.8608 at epoch 27
 MFCC-39 — best val. accuracy 0.8667 at epoch 33

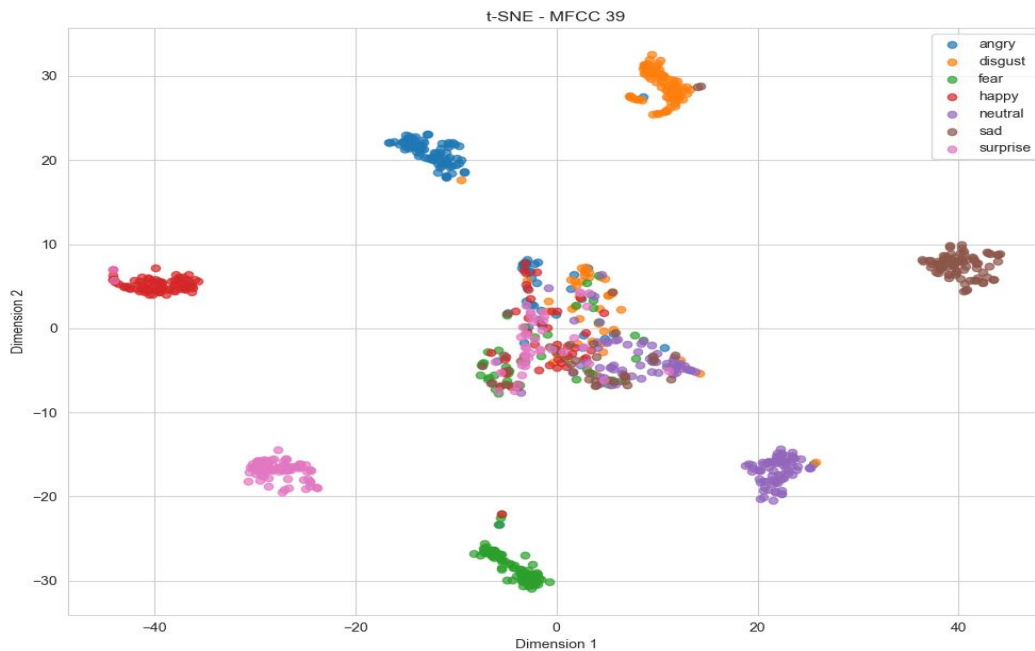


Figure no 4: MFCC-39 — 64-D Dense embedding

VI. DISCUSSION

The main purpose of this study is to understand how the number of MFCC coefficients affects the performance of a Stacked GRU model for speech emotion recognition.

MFCC-13 provides a compact input representation and may be suitable when computational efficiency is important. MFCC-26 provides an intermediate configuration, while MFCC-39 represents the speech signal using a larger number of coefficients. A larger feature representation may preserve additional acoustic information that helps distinguish emotions with similar speech characteristics. However, increasing the number of coefficients also increases the input dimensionality and may affect computational requirements. Therefore, the best configuration should be selected using experimental evidence rather than assuming that a larger feature set will always perform better.

The evaluation approach also considers prediction confidence. A model may classify a sample correctly but assign an unnecessarily high probability to its prediction. Calibration analysis helps identify this behaviour and provides a reason to investigate techniques such as temperature scaling. The t-SNE visualisations provide a way to inspect the learned feature representations. However, visual separation alone does not establish that a model will generalise successfully to new speakers or recording environments.

Several limitations should be considered when interpreting the findings. The study uses a relatively small combined dataset, and both source datasets contain controlled emotional speech. TESS includes only two female speakers, which limits speaker diversity. The experimental results may therefore differ when the model is tested on spontaneous speech, unfamiliar accents, other languages, or recordings containing background noise.

Future evaluation should include speaker-independent testing, additional speech datasets, noise robustness experiments, and measurements of inference time. These steps would provide stronger evidence of the model's suitability for practical applications.

VII. CONCLUSION

This study presents a deep-learning framework for Speech Emotion Recognition using sequential MFCC features and a Stacked GRU classifier. The project combines RAVDESS and TESS into a dataset of 4,240 audio samples covering seven emotion categories. Three MFCC configurations containing 13, 26, and 39 coefficients are evaluated using the same model architecture and training procedure. The evaluation framework includes classification metrics, confusion matrices, ROC-AUC, probability calibration curves, and t-SNE visualisations. The study provides a structured approach to comparing MFCC representations for GRU-based speech emotion classification. Future work can extend the framework through speaker-independent evaluation, testing under different noise conditions, real-time inference measurements, and validation on natural speech recordings. These improvements would help determine how well the proposed approach performs beyond the controlled benchmark datasets.

REFERENCES

1. S. R. Livingstone and F. A. Russo, "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*, vol. 13, no. 5, 2018.
2. K. Dupuis and M. K. Pichora-Fuller, "Toronto Emotional Speech Set (TESS)," 2010.
3. S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2227–2231.
4. K. Cho et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proceedings of the*

- Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.
5. F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
 6. M. McFee et al., “librosa: Audio and music signal analysis in Python,” in *Proceedings of the Python in Science Conference*, 2015, pp. 18–25.
 7. M. Abadi et al., “TensorFlow: A system for large-scale machine learning,” in *Proceedings of the USENIX Symposium on Operating Systems Design and Implementation*, 2016.
 8. L. van der Maaten and G. Hinton, “Visualizing data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.
 9. A. Niculescu-Mizil and R. Caruana, “Predicting good probabilities with supervised learning,” in *Proceedings of the International Conference on Machine Learning*, 2005, pp. 625–632.