Enhancing Speech Emotion Recognition using Long Short Term Memory Network

Under the Guidance of

Dr K. Sravanthi

Associate Professor

Presented by C-18 Batch

Peddi Ajay Kumar (Y21EC138)

Yericharla Sayomi (Y21EC196)

Ramavath Mallikharjuna Naik (Y21EC146)

Objectives

- To address the limitations of the traditional CNN-LSTM approach, which struggles to capture the complex temporal dependencies in speech.
- To incorporates MFCCs as supplementary features, initial normalization, standard scaling, data reshaping is need to do for TESS Dataset.
- Design a LSTM model to capture the sequential patterns and long-range temporal dependencies in the audio files of TESS dataset, which is crucial for accurate emotion recognition.
- To correctly identify the emotions being expressed in the speech signals of TESS dataset. This could include basic emotions like happiness, sadness, anger, fear, surprise, and disgust, neutral.
- To use the model in applications, like interactive voice assistants or call center analysis, the model needs to be efficient enough to process and classify emotions in near real-time.

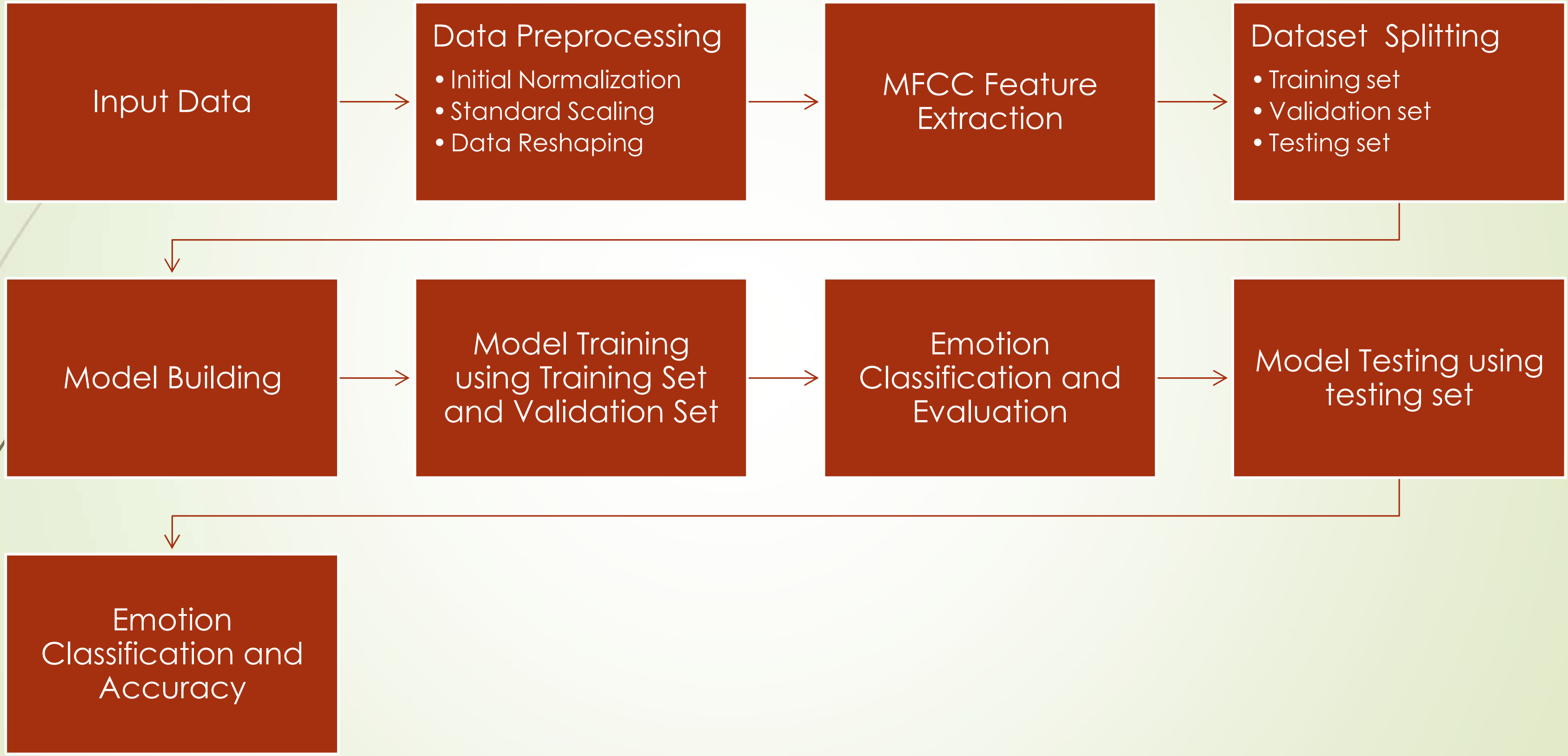
Problem Statement

- Due to the shortcomings of the previously used approach that combines Convolutional Neural Networks (CNN) with Long Short-Term Memory (LSTM) networks for SER tasks, it faces a difficult problem.
- Although this fusion method has demonstrated potential, it is hampered by LSTM cells' intrinsic restriction, which is that they only store the CNN's output for a single point in time.
- This feature does less well when it comes to identifying emotions since it is unable to capture the intricate temporal correlations found in time series data.
- To address this problem we would like to propose a design of LSTM model to capture the sequential patterns and long-range temporal dependencies in the audio files of TESS dataset, which is crucial for accurate emotion recognition.

Introduction

- Speech emotion recognition (SER) is the process of using machine learning to identify and analyze emotions in speech. It's a subdiscipline of affective computing that can help improve human-machine interactions. SER systems work by processing and classifying speech signals to detect emotions.
- In this project to examine the performance of the speech emotion recognition (SER) system, we used TESS dataset namely "Toronto Emotional Speech Set" which contains 2800 speech files for 7 emotions each emotion has 400 speech files by two female speakers labeled as (YAF and OAF).
- Long Short-Term Memory(LSTM) is well-suited for processing sequential data like speech. LSTMs enabling the network to effectively learn and retain long-term dependencies within the input sequences, which is crucial for recognizing patterns and meaning in speech.
- Additionally, we incorporate Mel-frequency cepstral coefficients (MFCCs) as supplementary features to enrich the representation of spoken words. MFCCs are known for their effectiveness in capturing the spectral characteristics of audio signals, which are vital for emotion recognition.
- To enhance SER system, we used libraries namely Librosa, TensorFlow, NumPy, pandas, matplotlib, seaborn, etc.

Flow Chart



Input Data

Data Preprocessing

- Initial Normalization
- Standard Scaling
- Data Reshaping

MFCC Feature Extraction

Dataset Splitting

- Training set
- Validation set
- Testing set

Model Building

Model Training using Training Set and Validation Set

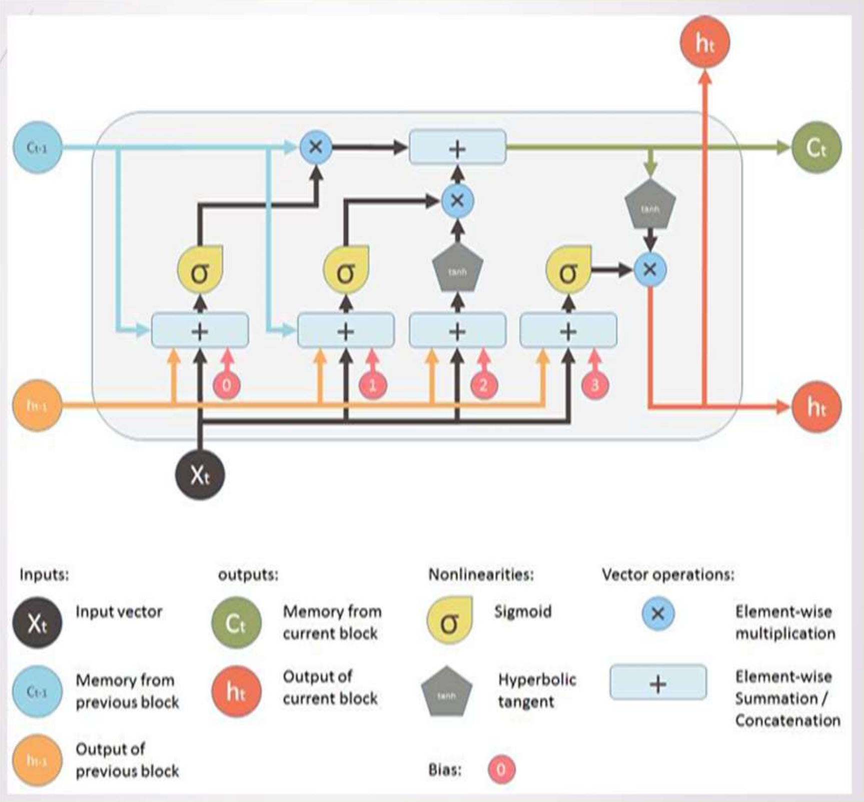
Emotion Classification and Evaluation

Model Testing using testing set

Emotion Classification and Accuracy

Model Architecture

- This architecture is designed to capture complex temporal dependencies an extract com plex patterns from the input speech data. LSTM networks are composed of memory cells and various gates that regulate the flow of information within the network.
- **Input Gate** : $i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$
- **Forget Gate** : $f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$
- **Candidate Cell State** : $\hat{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$
- **Cell State Update** : $c_t = f_t \cdot c_{t-1} + i_t \cdot \hat{c}_t$
- **Output Gate** : $o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$



h_t
C_{t-1}
C_t
σ
σ
σ
tanh
tanh
h_{t-1}
h_t
0
1
2
3
X_t
Inputs:
outputs:
Nonlinearities:
Vector operations:
X_t Input vector
C_t Memory from current block
σ Sigmoid
× Element-wise multiplication
C_{t-1} Memory from previous block
h_t Output of current block
tanh Hyperbolic tangent
+ Element-wise Summation / Concatenation
h_{t-1} Output of previous block
Bias: 0

Model Architecture

The above table show a summary implementation of the deep learning model.

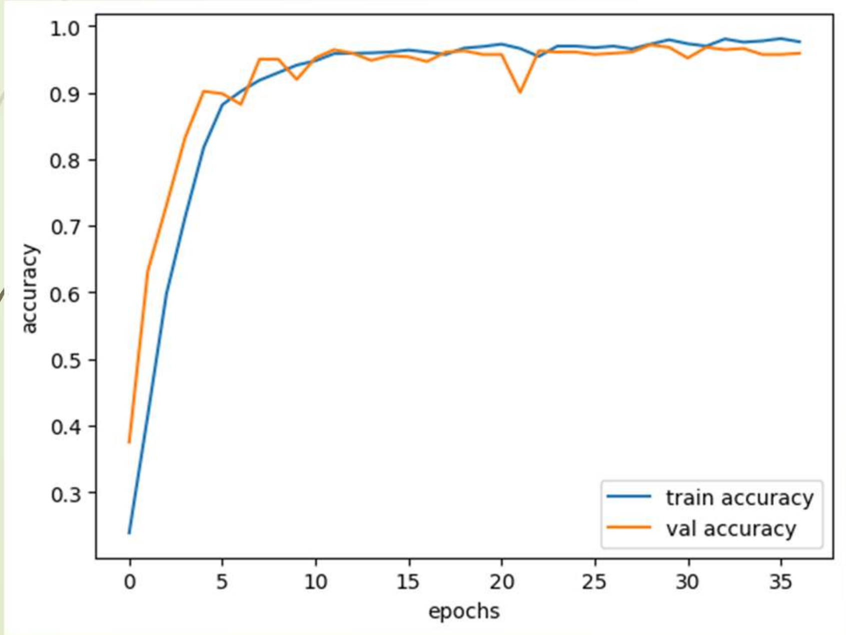
The first column shows the layers and layer types, the second column shows output shape which is the number of units or neurons for a particular layer and the third column shows the parameters per layer.

The model has a total of 527,751 trainable parameters, indicating a relatively complex model capable of learning intricate patterns in the data.

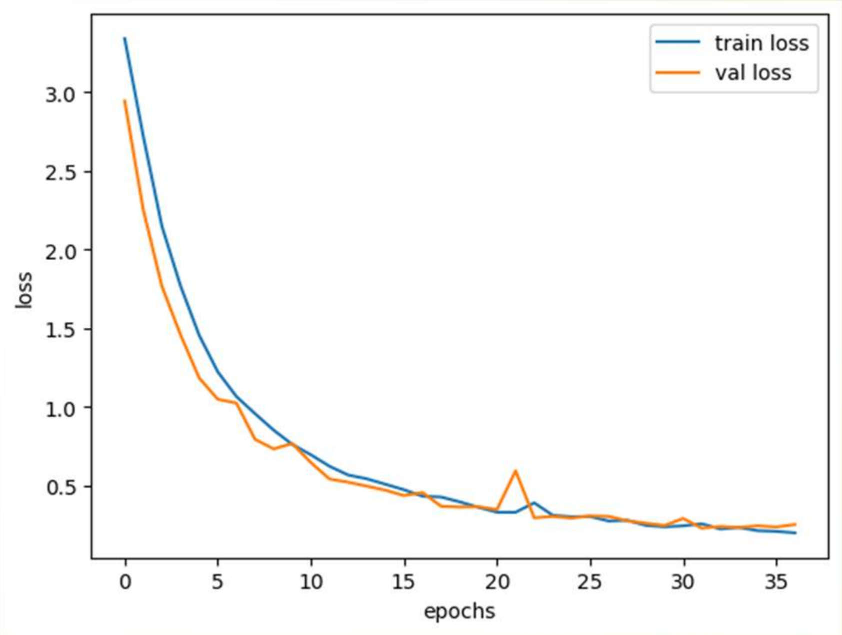
Layer (type)	Output Shape	Parameters
lstm (LSTM)	(None, 40 256)	264,192
dropout (Droput)	(None ,40 ,256)	0
lstm_1 (LSTM)	(None ,40, 128)	197, 120
dropout_1(Dropout)	(None, 40, 128)	0
lstm_2 (LSTM)	(None, 64)	49, 408
dropout_2(Dropout)	(None, 64)	0
dense (Dense)	(None, 128)	8, 320
dropout_3(Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 64)	8, 256
dropout_4(Dropout)	(None, 64)	0
dense_2 (Dense)	(None, 7)	455

Results

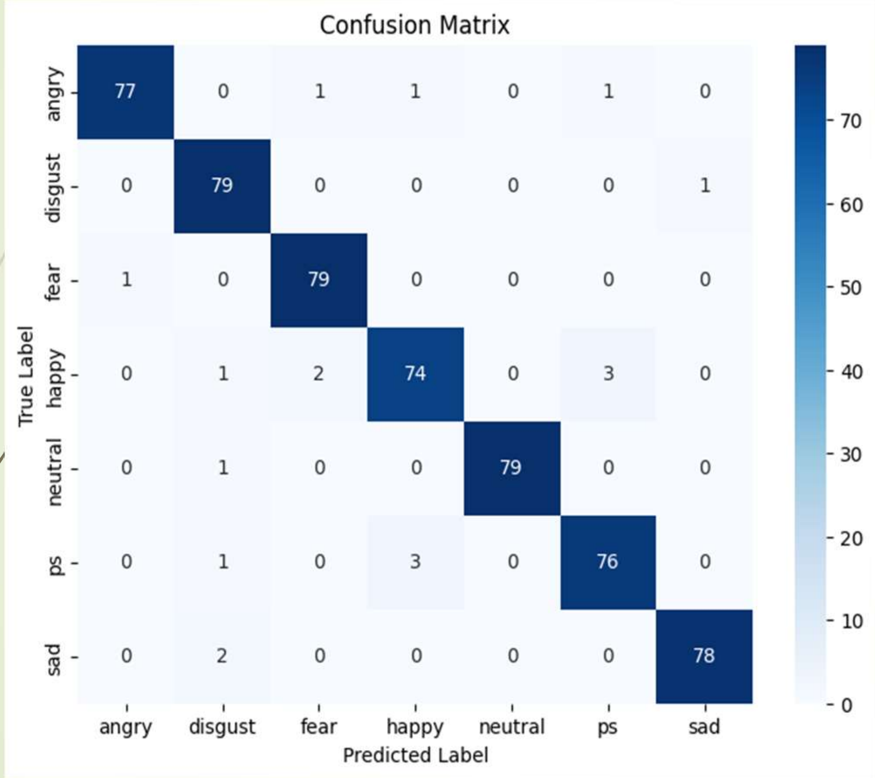
Training Accuracy vs Validation Accuracy



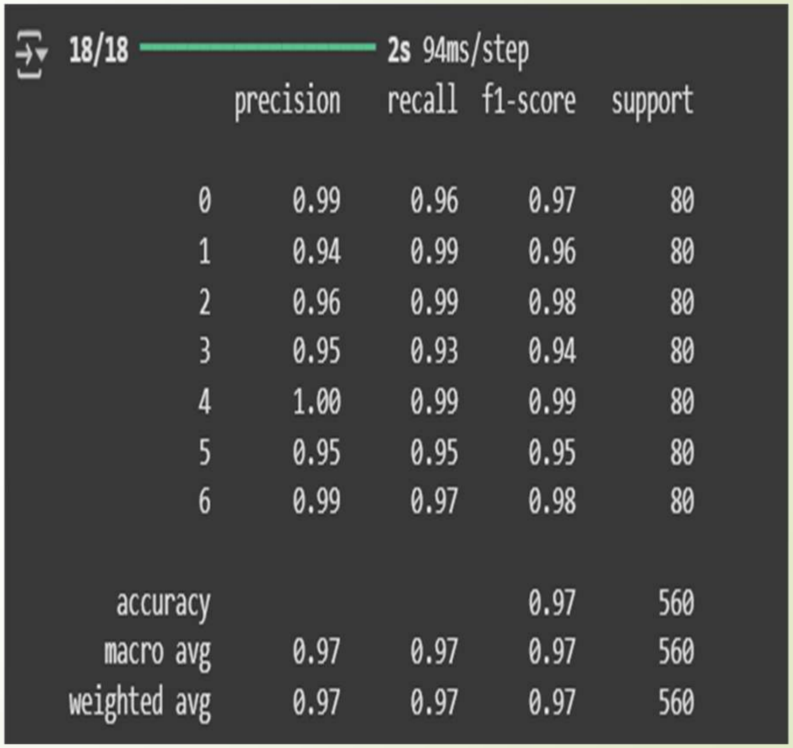
Train Loss vs Validation Loss



Emotion Classification for Validation Set



True Label \ Predicted Label	angry	disgust	fear	happy	neutral	ps	sad
angry	77	0	1	1	0	1	0
disgust	0	79	0	0	0	0	1
fear	1	0	79	0	0	0	0
happy	0	1	2	74	0	3	0
neutral	0	1	0	0	79	0	0
ps	0	1	0	3	0	76	0
sad	0	2	0	0	0	0	78

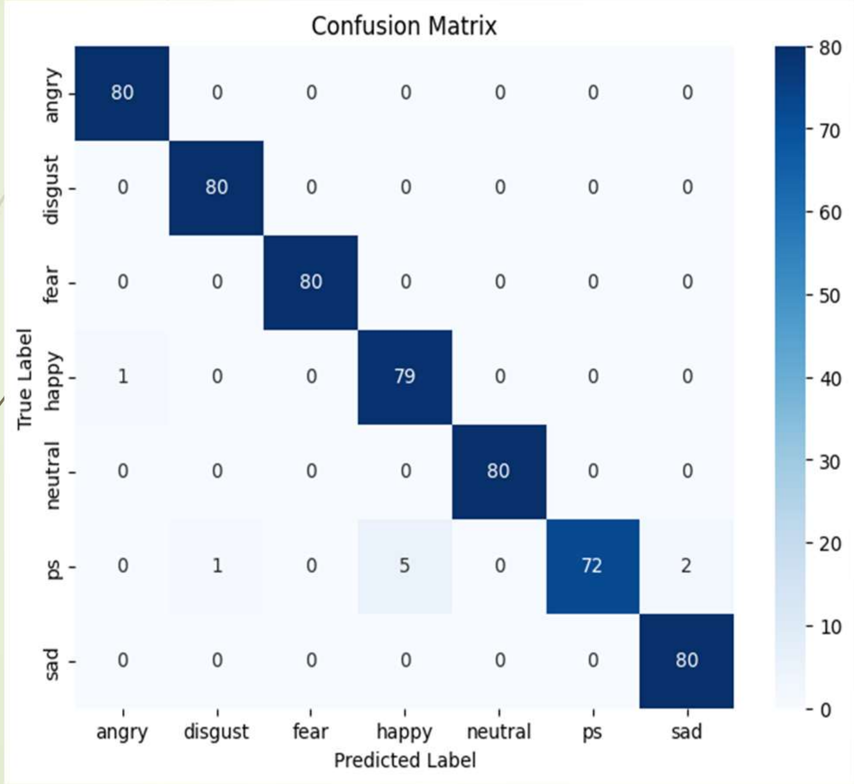


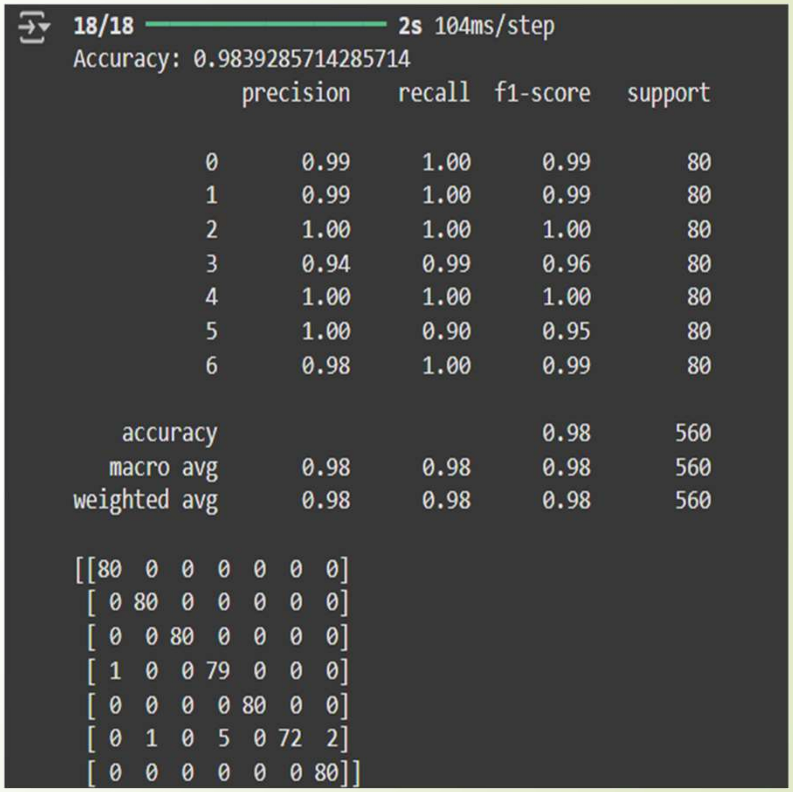
```
18/18 ━━━━━━━━━━━━━━━━━━━━ 2s 94ms/step
              precision    recall  f1-score   support

           0       0.99      0.96      0.97        80
           1       0.94      0.99      0.96        80
           2       0.96      0.99      0.98        80
           3       0.95      0.93      0.94        80
           4       1.00      0.99      0.99        80
           5       0.95      0.95      0.95        80
           6       0.99      0.97      0.98        80

    accuracy                           0.97       560
   macro avg       0.97      0.97      0.97       560
weighted avg       0.97      0.97      0.97       560
```

Emotion Classification for Testing Set





```
18/18 ━━━━━━━━━━━━━━━━━━━━ 2s 104ms/step
Accuracy: 0.9839285714285714
              precision    recall  f1-score   support

           0       0.99      1.00      0.99        80
           1       0.99      1.00      0.99        80
           2       1.00      1.00      1.00        80
           3       0.94      0.99      0.96        80
           4       1.00      1.00      1.00        80
           5       1.00      0.90      0.95        80
           6       0.98      1.00      0.99        80

    accuracy                           0.98       560
   macro avg       0.98      0.98      0.98       560
weighted avg       0.98      0.98      0.98       560

[[80  0  0  0  0  0  0]
 [ 0 80  0  0  0  0  0]
 [ 0  0 80  0  0  0  0]
 [ 1  0  0 79  0  0  0]
 [ 0  0  0  0 80  0  0]
 [ 0  1  0  5  0 72  2]
 [ 0  0  0  0  0  0 80]]
```

Comparison with Other Models

Reference	Features	Dataset	Model	Accuracy	Comments
Muralidharan et al	Vocal Tract Prosodic Non-Linear	RAVDESS + TESS + SAVEE + CREMA-D	CNN+LSTM	72.66%	Contribute to provide better models for emotion detection for a bot, for interacting with humans.
Beena Salian et al	Mel Spectrograms	RAVDESS + TESS + SAVEE	CNN+LSTM	89.26%	The optimizer for this model is Stochastic Gradient Descent (SGD).
Passricha et al	MFCC	TESS	CNN + Bi-LSTM	86.43%	Non-linearity with dropout idea shows 5.8% and 10% relative decrease in WER over the CNN and DNN systems, respectively.
J. Parry et al	Mel Filter Bank Coefficients	Cross-Corpus	CNN LSTM CNN+LSTM	53.35%	Performed on Cross-Corpus and TESS corpus was reserved as a out-of-domain test set.

Comparison with Other Models

Reference	Features	Dataset	Model	Accuracy	Comments
Neha Prerna Tigga et al	MFCC	SAVEE + RAVDESS + TESS	Hybrid CNN+LSTM	76.16%	A CuDNNLSTM layer was applied to LSTM layer for faster Implementation using GPU Processing.
Faith Sengul et al	MFCC	TESS	LSTM	90%	Emotion Classification based on Speaker gender.
Testcase on EMO-DB	MFCC	EMO-DB	Bi-LSTM	90%	The sparse categorical cross-entropy loss function is used.
Testcase on RAVDESS	MFCC	RAVDESS	Single Layer LSTM	54%	The MFCC features are saved in xlsx file.
Proposed Model	MFCC	TESS	3-Layer Stacked LSTM	98.39%	The categorical cross-entropy loss function is used.

Conclusion

In this study, we developed a deep learning model for speech emotion recognition using three-layer stacked LSTM architecture. Through extensive data preprocessing, feature extraction and model optimization, our model outperformed the benchmark of hybrid CNN+LSTM model.

While our proposed model has demonstrated promising results on TESS dataset, it's performance on smaller datasets are not good, this highlights an area of investigation. This limitation suggests potential challenges in generalizing to datasets with limited data diversity and size.

By addressing these aspects in future research, we aim to enhance the LSTM's model's performance on smaller datasets and bridge the gap between laboratory setting and real-world applications. This would facilitate the development of more robust and reliable speech emotion recognition systems for diverse domains.

Thank You