Evaluating Speech Emotion Recognition through the lens of CNN & LSTM Deep Learning Models

1st Daniel F.O. Onah
Department of Information Studies
University College London
London, United Kingdom
d.onah@ucl.ac.uk

2nd Asia Ibrahim
Department of Information Studies
University College London
London, United Kingdom
asia.ibrahim.16@ucl.ac.uk

***Abstract*—Speech Emotion Recognition (SER) is a fascinating area of research in machine learning. Researchers have been exploring different techniques to improve this field including using deep learning models, feature extraction methods and transfer strategies to improve the accuracy and robustness of SER models. The advancements in SER not only contribute to the field of artificial intelligence but also have the potential to enhance our understanding of human emotions and improve communication between humans and machines.**

***Index Terms*—speech emotion recognition, CNN, LSTM, deep learning**

I. Introduction

The overall objective and motivation in this research is to evaluate the effectiveness of two machine learning algorithms for SER [4]. Using a Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) networks as the deep learning models [1]. The study aims to accurately predict emotional sentiments in audio speech samples. The research experiment includes a Conv1D layer, for the audio branch, to extract relevant features from the audio input. Each Conv1D layer is followed by a MaxPooling layer to reduce the dimensionality of the dataset.

A. Research Questions

The research questions in this study were as follows:

- How effective are CNN and LSTM networks in recognizing emotions from speech data?
- What is the impact of using Mel Frequency Cepstral Coefficients (MFCCs) for feature extraction on the performance of emotion recognition using CNN and LSTM algorithms?
- To what extent do the results of the CNN and LSTM models align with the objective of improving the accuracy of Speech Emotion Recognition (SER) systems?

II. Methodology

The data used in this study is from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). The LSTM model was compiled as a sequential neural network with three LSTM layers (70, 50, and 60 units), followed by dense layers (48 units with ReLU activation, 14 units) [7]. The model underwent 50 epochs of training with a batch size of 28. The CNN model was also compiled sequentially and consisted of Conv1D layers with increasing filters (1024, 512, 256, 128), each followed by max-pooling. A GlobalAveragePooling1D layer was applied, followed by dense layers (64 units, ReLU, 0.2 dropout, 14 units with softmax) (see Figure 1). It underwent 100 epochs of training with a batch size of 30. Both CNN & LSTM models were compiled using Adam optimizer and also the computation of the mean squared error loss and accuracy metric were performed within the models. Figure 1 illustrates the model architecture for the study.

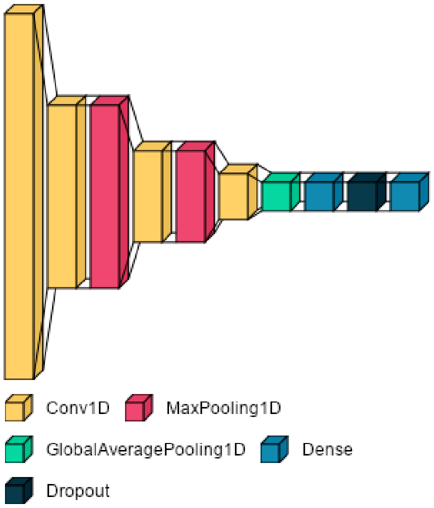


Fig. 1. CNN Model Architecture

A. Tuning Hyperparameter

In our model landscape architecture, the hyperparameters are in threefold: batch size, learning rate, and the number of iterative epochs. Each of these parameters contributed adequately to the efficiency of the model while taking into consideration their independent functions to improve the performance of the model. The learning rate of 0.001 contributed sufficiently to the accuracy of the models. The rate ensures improved training with the epochs and the batch size. The model optimization was performed using Adam optimizer with the learning rate. Adam optimizer is an optimization algorithm, useful for handling sparse gradients on noisy data point [5] [6].

III. Experiment and Results

LSTM and CNN has been shown to be highly effective for sound and signal processing tasks [1]. It has been able to identify the intricate patterns in speech data that correspond to various emotions, audio data with noise and produces extremely important accurate prediction results. The Figures below illustrates two audio data, Figure 2 is a normal audio without noise and Figure 3 is an audio with additional noise layer that was introduced during the experiment. Further experiment conducted using 100 epochs in a batch size of 30, shows improved outcome, CNN has produced a training accuracy of 97% and a test accuracy of 64%. The initial experiment conducted using 50 epochs when CNN model was used in conjunction with LSTM was able to produce a training accuracy of 68% and test accuracy of 53%. The CNN+LSTM model for the training set swiftly converges to optimal performance, with its rapidly decreasing loss complementing its soaring accuracy level as illustrated in Figure 6 and Figure 7.

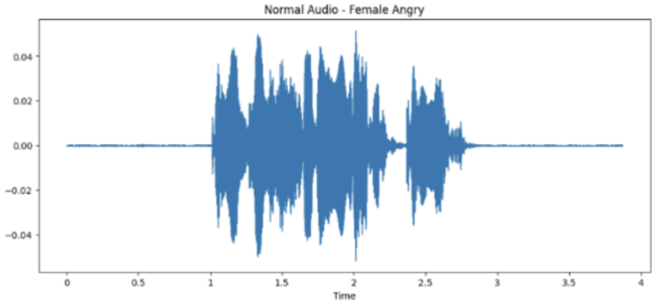


Fig. 2. Normal audio data

The normal audio data is further mashed with noise and passed through the deep-learning functions to understand and evaluate the dataset from a different dimension.

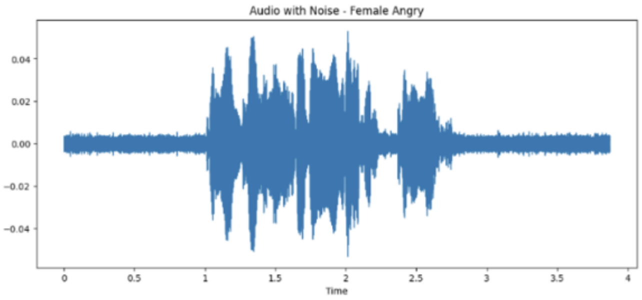


Fig. 3. Audio with noise

A. Mel Spectrogram and MFCC results

Mel Spectrogram and MFCC graphs were used to visualise audio speech patterns and allow for comparison between male and female audio samples [9]. Figure 4 shows Mel Spectrogram of female happy emotions and Figure 5 shows male happy emotions.

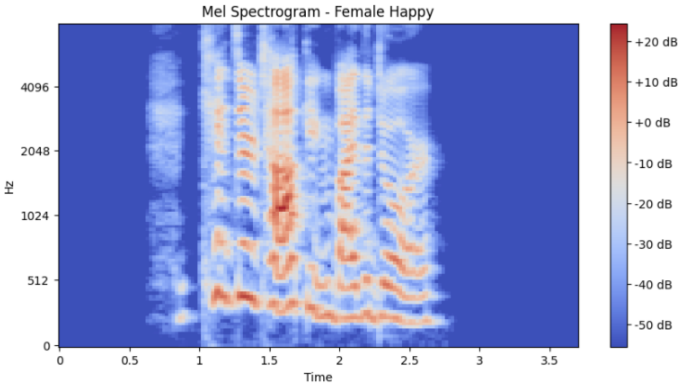


Fig. 4. Mel Spectrogram female 'happy' emotion

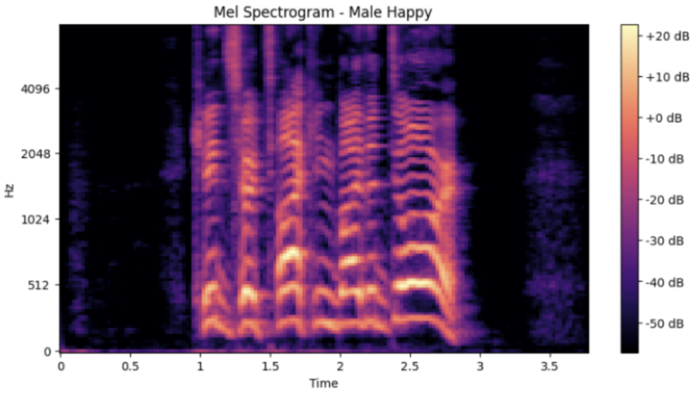


Fig. 5. Mel Spectrogram male 'happy' emotion

B. Measuring metrics

Accuracy and confusion matrix were chosen as the measuring metrics to evaluate the performance of the models (see Figure 6, Figure 7 and Figure 8). Confusion matrix was used to elicitate the performance of the model. Confusion matrix is an **N** by **N** matrix, where *N* is the number of classes. Each row and column of the matrix represents instances of the predicted class and actual class respectively [8]. To ensure a thorough evaluation, additional metrics such as precision, recall, and F1-score were also considered to garner a more holistic understanding of the model's capabilities and performance.

C. MFCC comparison graph

This comparison shows the difference in male and female speech frequency level when displaying surprise emotions. Using MFCC, we were able to identify the acoustic pitch level [3] or strength within the emotional tones of both male and female surprise emotions as illustrated in Figure 9.

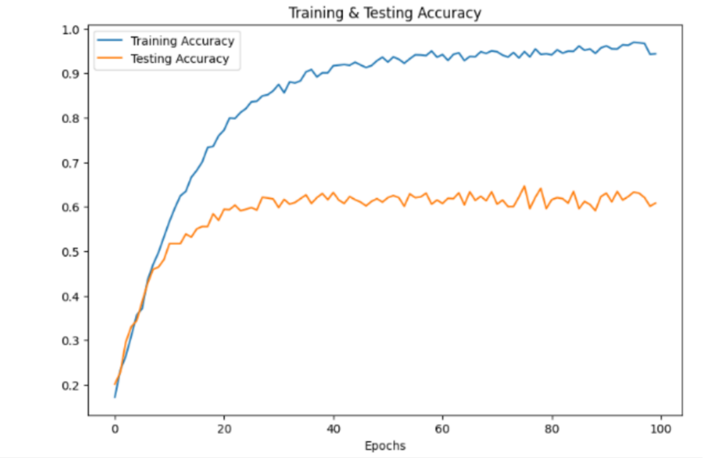


Fig. 6. Training and testing accuracy

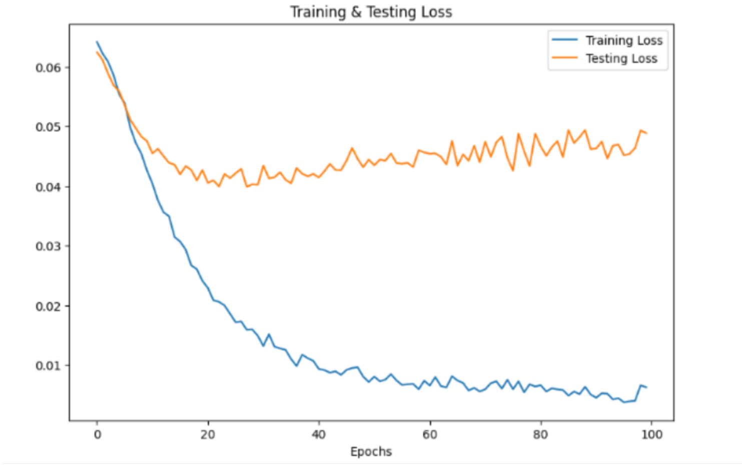


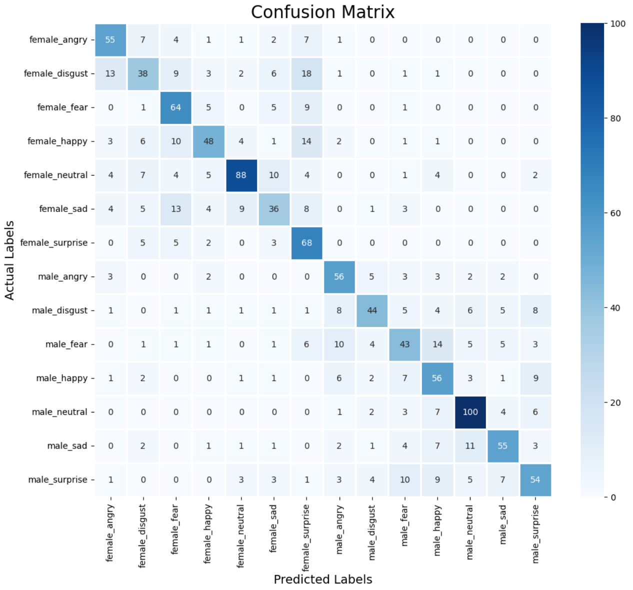
Fig. 7. Training and testing loss

IV. Conclusion

The divergent behaviours between CNN and LSTM independent models and the CNN+LSTM models also hint at their varying capabilities in tackling audio datasets. Earlier experiment results show that the deep learning models outperform both traditional machine learning algorithms such as Decision Tree and KNN models individually, with CNN achieving a better accuracy result and improving precision, recall, and F-score [2]. The findings in this research experiment suggest that deep learning techniques such as CNN and LSTM either as standalone models or combined can be effective tools for predicting SER, and that deep learning models can further enhance their predictive power in speech emotion recognition tasks.

References

[1] Fayek, H. M., Lech, M., and Cavedon, L. Evaluating deep learning architectures for speech emotion recognition. *Neural Networks 92* (2017), 60–68.



	female_angry	female_disgust	female_fear	female_happy	female_neutral	female_sad	female_surprise	male_angry	male_disgust	male_fear	male_happy	male_neutral	male_sad	male_surprise
female_angry	55	7	4	1	1	2	7	1	0	0	0	0	0	0
female_disgust	13	38	9	3	2	6	18	1	0	1	1	0	0	0
female_fear	0	1	64	5	0	5	9	0	0	1	0	0	0	0
female_happy	3	6	10	48	4	1	14	2	0	1	1	0	0	0
female_neutral	4	7	4	5	88	10	4	0	0	1	4	0	0	2
female_sad	4	5	13	4	9	36	8	0	1	3	0	0	0	0
female_surprise	0	5	5	2	0	3	68	0	0	0	0	0	0	0
male_angry	3	0	0	2	0	0	0	56	5	3	3	2	2	0
male_disgust	1	0	1	1	1	1	1	8	44	5	4	6	5	8
male_fear	0	1	1	1	0	1	6	10	4	43	14	5	5	3
male_happy	1	2	0	0	1	1	0	6	2	7	56	3	1	9
male_neutral	0	0	0	0	0	0	0	1	2	3	7	100	4	6
male_sad	0	2	0	1	1	1	0	2	1	4	7	11	55	3
male_surprise	1	0	0	0	3	3	1	3	4	10	9	5	7	54

Fig. 8. Confusion Matrix SER Classification

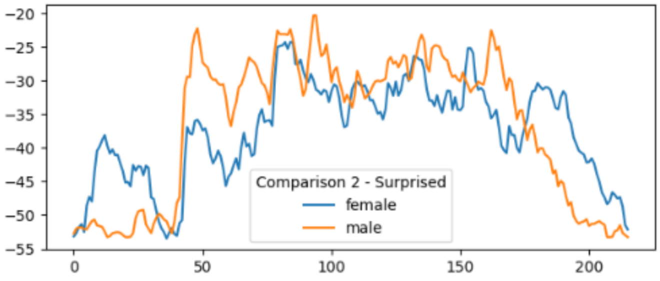


Fig. 9. MFCC graph comparing male and female ‘surprise’

[2] Huang, K.-Y., Wu, C.-H., Hong, Q.-B., Su, M.-H., and Zeng, Y.-R. Speech emotion recognition using convolutional neural network with audio word-based embedding. In *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)* (2018), IEEE, pp. 265–269.

[3] Jin, Q., Li, C., Chen, S., and Wu, H. Speech emotion recognition with acoustic and lexical features. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (2015), IEEE, pp. 4749–4753.

[4] Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., Mahjoub, M. A., and Cleder, C. Automatic speech emotion recognition using machine learning, 2019.

[5] Kingma, D. P., and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[6] Liu, M., Zhang, W., Orabona, F., and Yang, T. Adam ˆ+: A stochastic method with adaptive variance reduction. *arXiv preprint arXiv:2011.11985* (2020).

[7] Wang, J., Xue, M., Culhane, R., Diao, E., Ding, J., and Tarokh, V. Speech emotion recognition with dual-sequence lstm architecture. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2020), IEEE, pp. 6474–6478.

[8] Žgank, A., Horvat, B., and Kačič, Z. Data-driven generation of phonetic broad classes, based on phoneme confusion matrix similarity. *Speech Communication 47*, 3 (2005), 379–393.

[9] Zhang, B., Leitner, J., and Thornton, S. Audio recognition using mel spectrograms and convolution neural networks. *Noiselab University of California: San Diego, CA, USA* (2019).