Speech Emotion Recognition Using LSTM Neural Network

Hesham Elshafey

The Idea:

Use LSTM model on extracted MFCC features to predict the emotional state of speakers

The Goal:

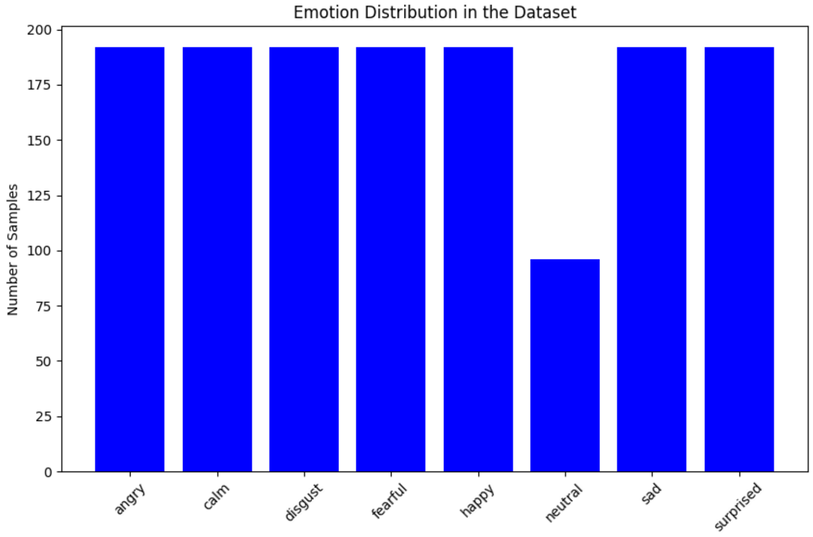
>70% prediction accuracy of eight emotional classes

Literature

- "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)" by Livingstone & Russo is licensed under CC BY-NA-SC 4.0.
- H. Ma, J. Wang, H. Lin, B. Zhang, Y. Zhang and B. Xu, "A Transformer-Based Model With Self-Distillation for Multimodal Emotion Recognition in Conversations," in IEEE Transactions on Multimedia, vol. 26, pp. 776-788, 2024, doi: 10.1109/TMM.2023.3271019.
- S. E. Eskimez, K. Imade, N. Yang, M. Sturge-Apple, Z. Duan and W. Heinzelman, "Emotion classification: How does an automated system compare to Naive human coders?," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 2016, pp. 2274-2278, doi: 10.1109/ICASSP.2016.7472082.
- Tzinis, E., Paraskevopoulos, G., Baziotis, C., & Potamianos, A. (2018). Integrating recurrence dynamics for speech emotion recognition. *Proceedings of Interspeech 2018*, 927–931.

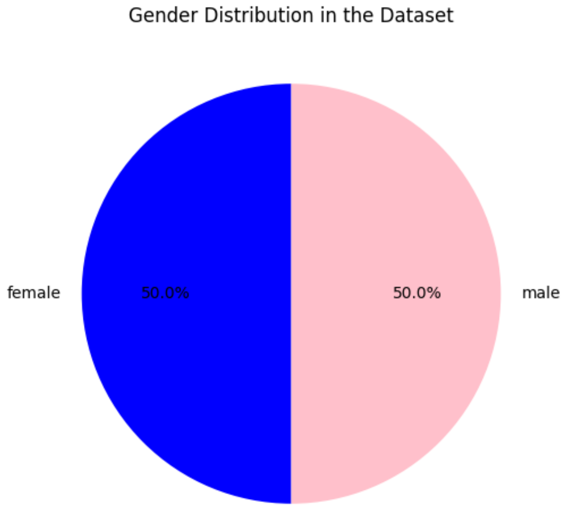
The dataset: RAVDESS

- Includes eight emotional states.
- 24 actors, each with many audio files of varying emotional states and intensity.
- Actors repeat the same sentence, ruling out “context” component.



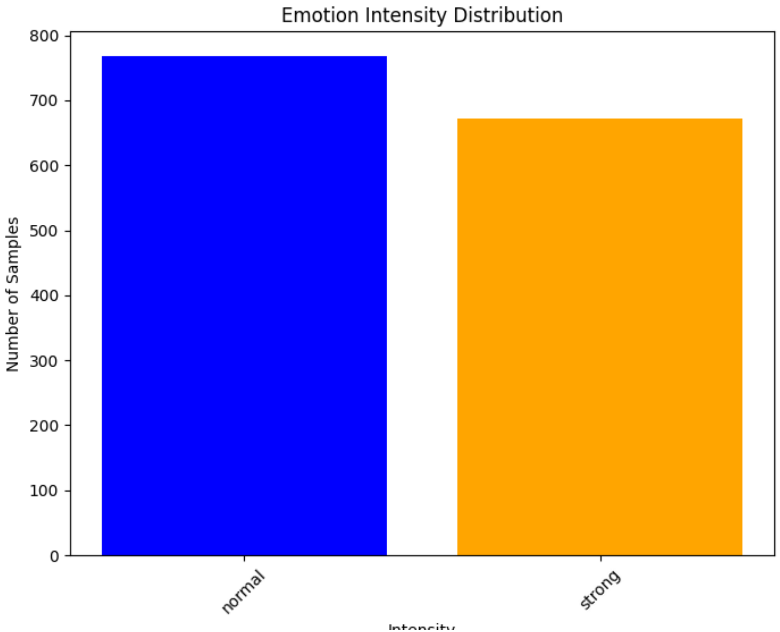
The dataset: RAVDESS

- Equal gender distribution, ruling out gender emotional inference from speech biases.



The dataset: RAVDESS

- Emotion is giving an intensity classification, with two classes.
- Helps with asserting the emotional inference.



The dataset: RAVDESS

- No CSV file! file names encoding the classification.

Filename example: 03-01-06-01-02-01-12.wav

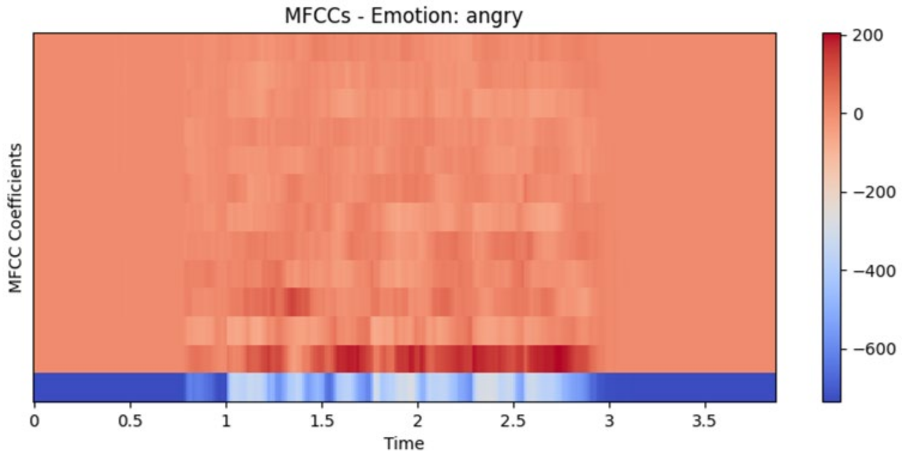
1. Audio-only (03)
2. Speech (01)
3. Fearful (06)
4. Normal intensity (01)
5. Statement "dogs" (02)
6. 1st Repetition (01)
7. 12th Actor (12)

Female, as the actor ID number is even.

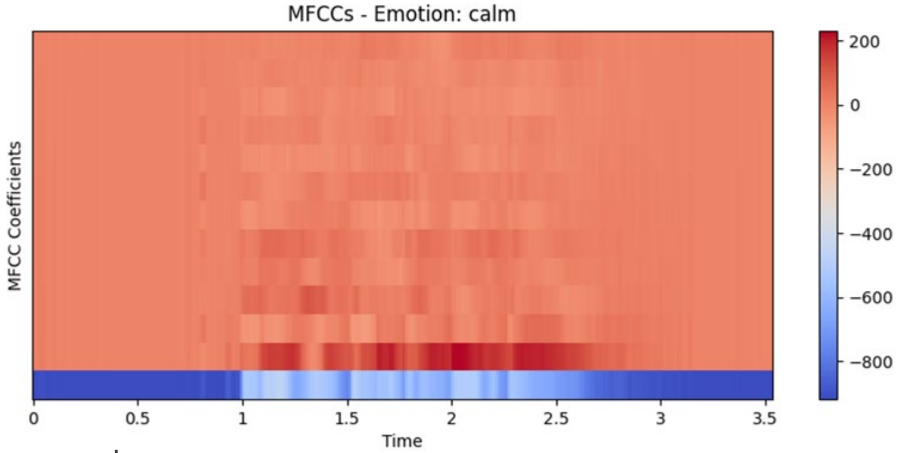
Why MFCC features?

- Provides robust representation of frequency content of speech that mimics human speech perception by capturing both frequency and amplitude.
- Emotions like anger and surprise are linked to discontinuous changes in frequency, with higher energy bursts apparent in the peaks of higher order MFCC coefficients.
- Emotions like sadness and calmness exhibit smoother frequency transitions with energy mostly in the lower frequencies domain. It manifests in the first few MFCC coefficients.

Why MFCC features?

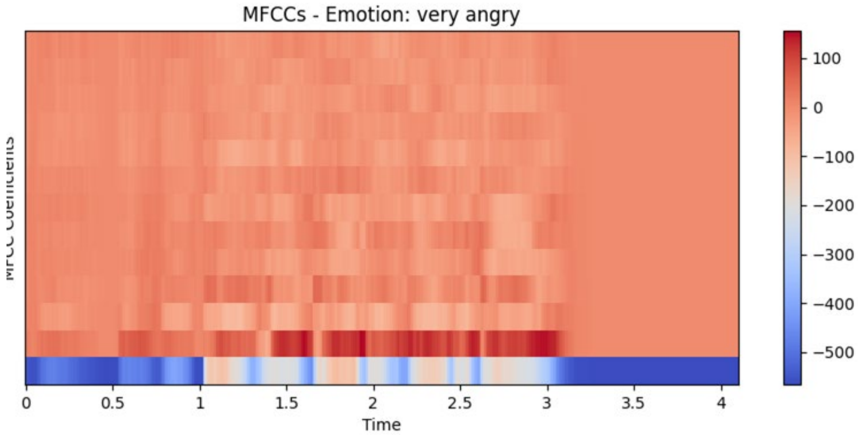


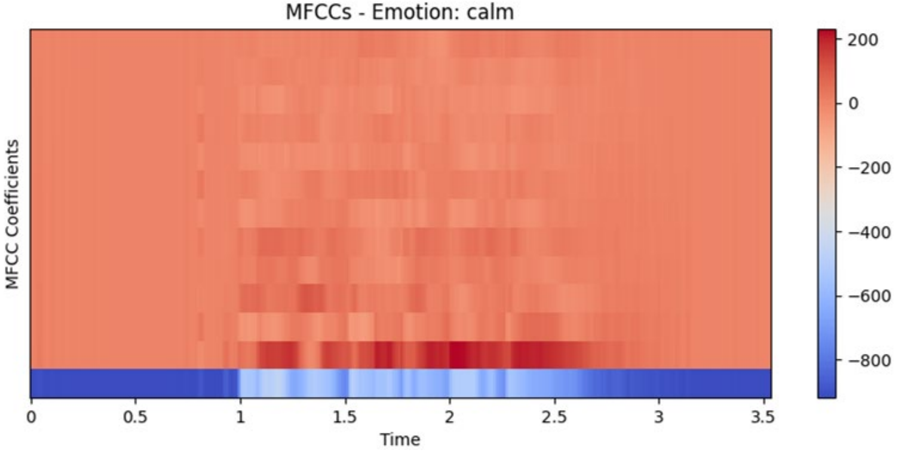
- Higher energy .
- Higher intensity indicating by the red regions.
- Higher order MFCC coefficients (about 9 bins).
- Sharper MFCC coefficients transitions.
- Sharper intensity peaks.



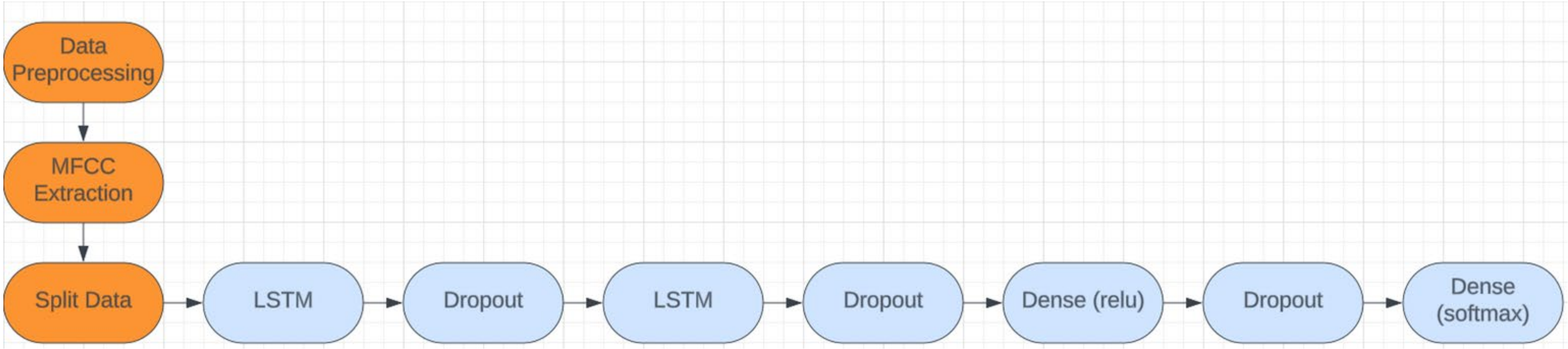
- Lower energy .
- Lower intensity indicating by the blue/orange regions.
- Lower order MFCC coefficients (about 5 bins).
- Smoother MFCC coefficients transitions.
- Smoother intensity peaks.

Why MFCC features?



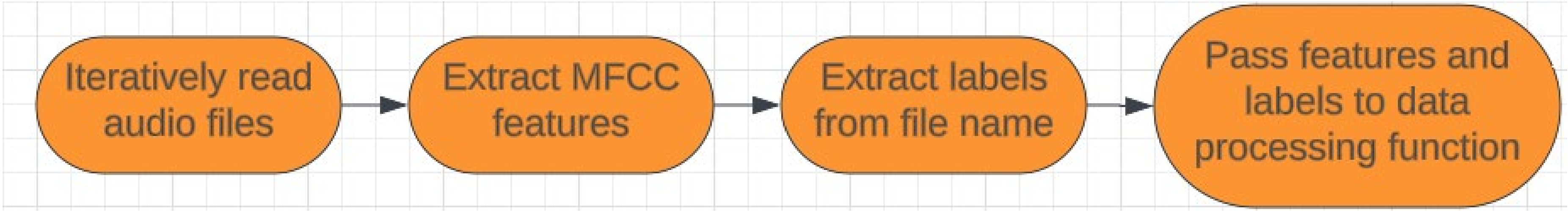


Model General Architecture (Rough)



- Adam Optimizer
- Spare Categorical cosentry
- Metrics = accuracy
- Early stopping

Data Preparation

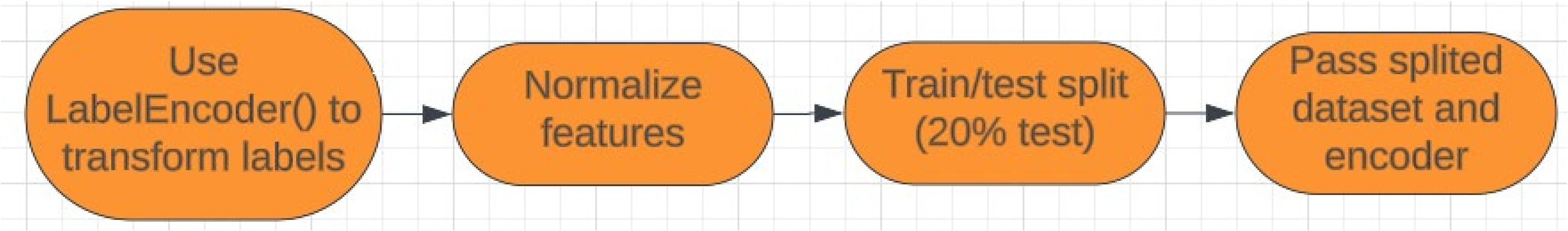


Filename example: 03-01-06-01-02-01-12.wav

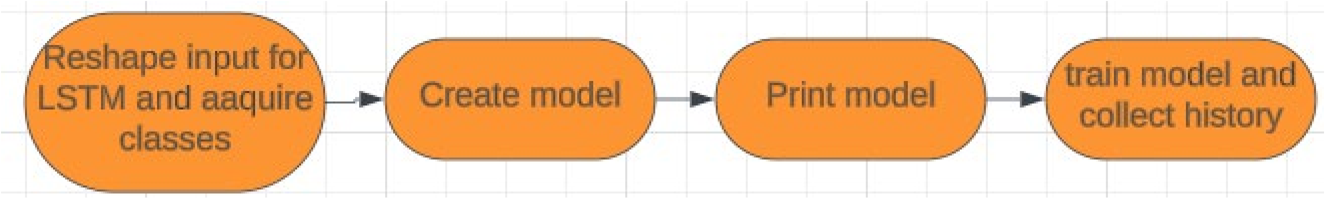
1. Audio-only (03)
2. Speech (01)
3. Fearful (06)
4. Normal intensity (01)
5. Statement "dogs" (02)
6. 1st Repetition (01)
7. 12th Actor (12)

Female, as the actor ID number is even.

Data Processing

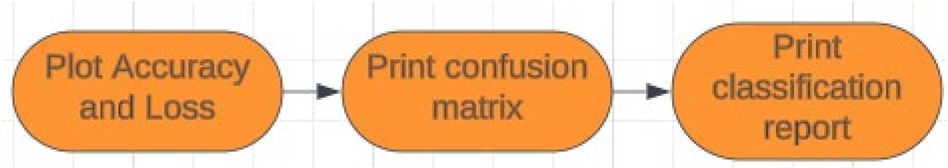


Training



Reshape input for LSTM and aaquire classes
Create model
Print model
train model and collect history

Evaluation



Plot Accuracy and Loss
Print confusion matrix
Print classification report

First model: Architecture

- Batch size = 32
- Epochs = 50
- Patience = 5

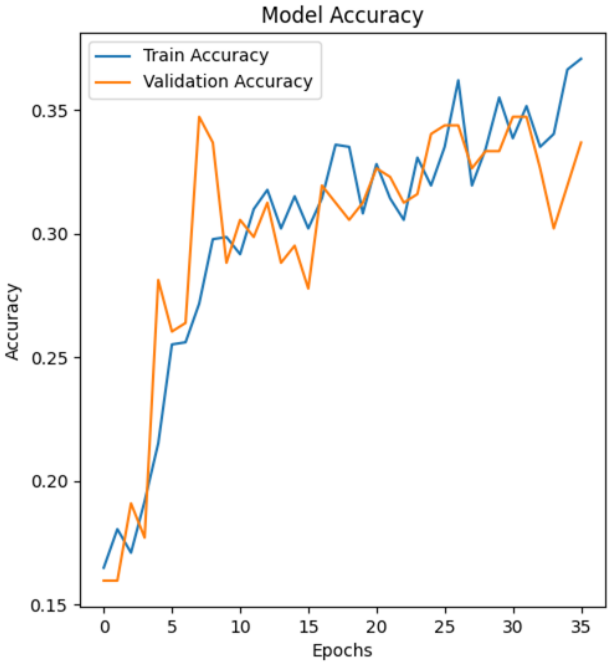
Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 13, 128)	66,560
dropout (Dropout)	(None, 13, 128)	0
lstm_1 (LSTM)	(None, 64)	49,408
dropout_1 (Dropout)	(None, 64)	0
dense (Dense)	(None, 128)	8,320
dropout_2 (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 8)	1,032

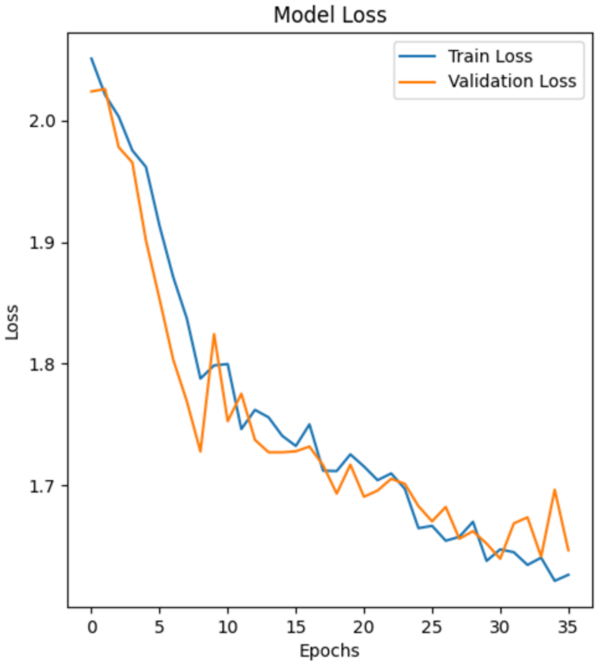
Total params: 125,320 (489.53 KB)

Trainable params: 125,320 (489.53 KB)

Non-trainable params: 0 (0.00 B)

First model: Evaluation

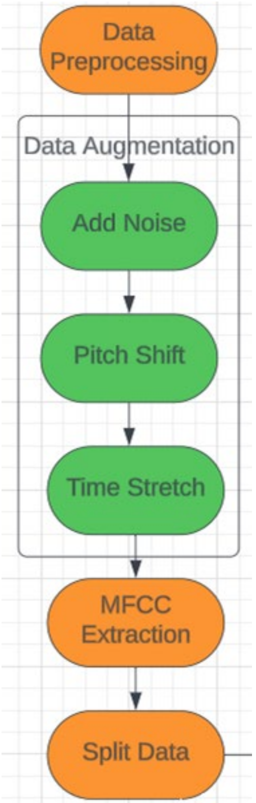




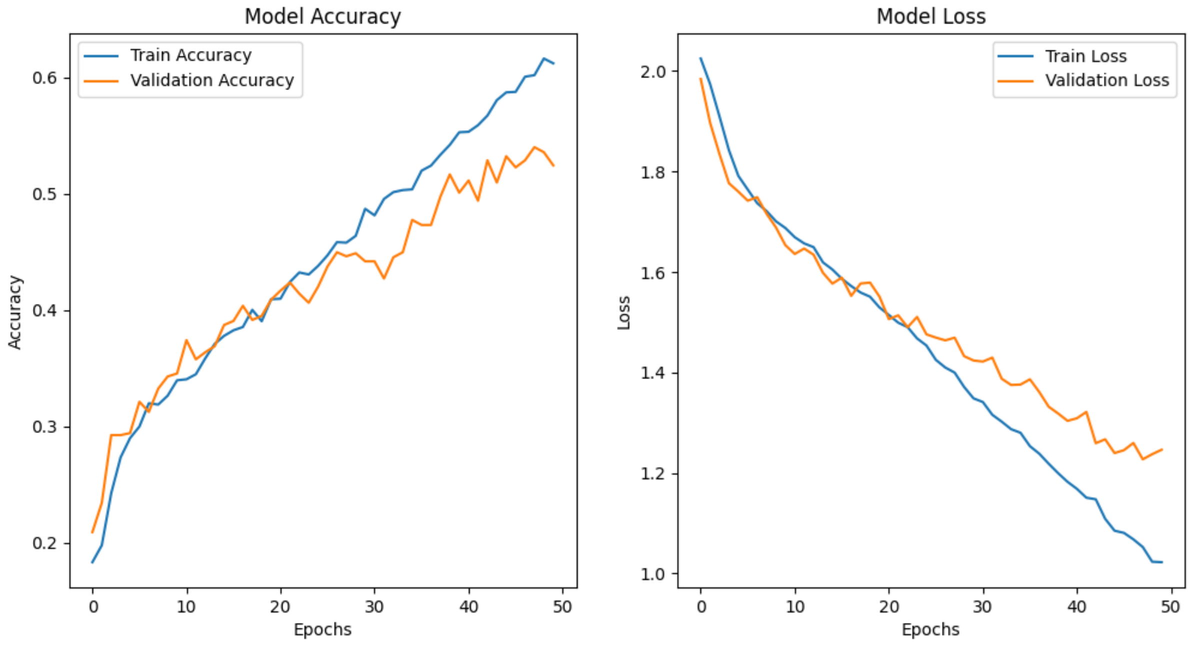
Pretty far from goal :(

Model Enhancement Attempt 1: Data Augmentation

- Diversify data by creating three variants of each file: Noisy, Pitched Shifted, and Time Stretched versions
- More data through another dataset would achieve the same result, though it is hard to find another data set with the same labels and features
- Added in the data preparation function



Model Enhancement Attempt 1: Data Augmentation - Results



Much better, but still not at the goal %!

Model Enhancement Attempt 2: Model Architecture and Training Changes

- Batch size = 32
- Epochs = **200**
- Patience = **10**

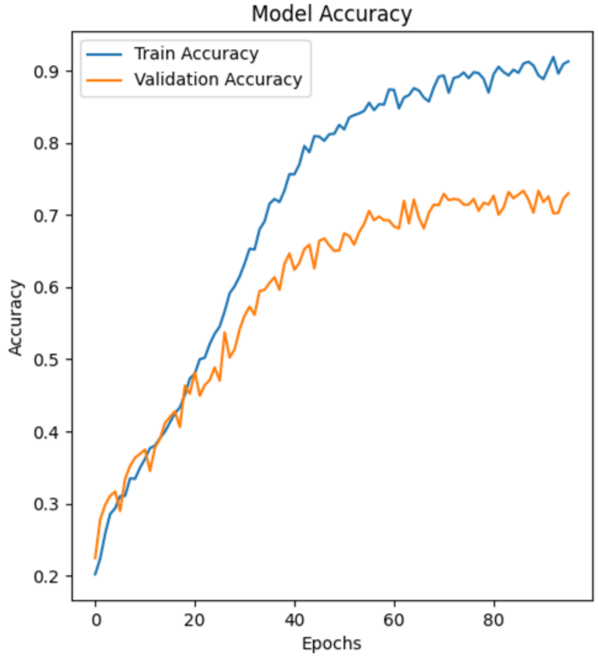
Layer (type)	Output Shape	Param #
lstm_2 (LSTM)	(None, 13, 256)	264,192
dropout_3 (Dropout)	(None, 13, 256)	0
lstm_3 (LSTM)	(None, 64)	82,176
dropout_4 (Dropout)	(None, 64)	0
dense_2 (Dense)	(None, 128)	8,320
dropout_5 (Dropout)	(None, 128)	0
dense_3 (Dense)	(None, 8)	1,032

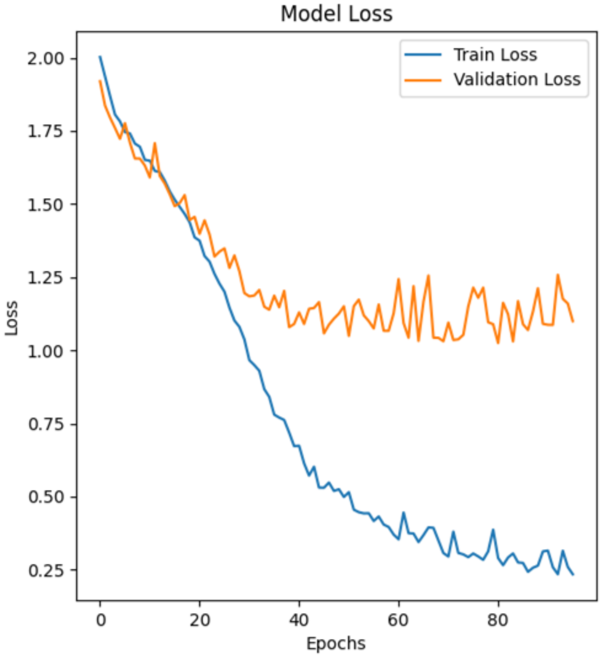
Total params: 355,720 (1.36 MB)

Trainable params: 355,720 (1.36 MB)

Non-trainable params: 0 (0.00 B)

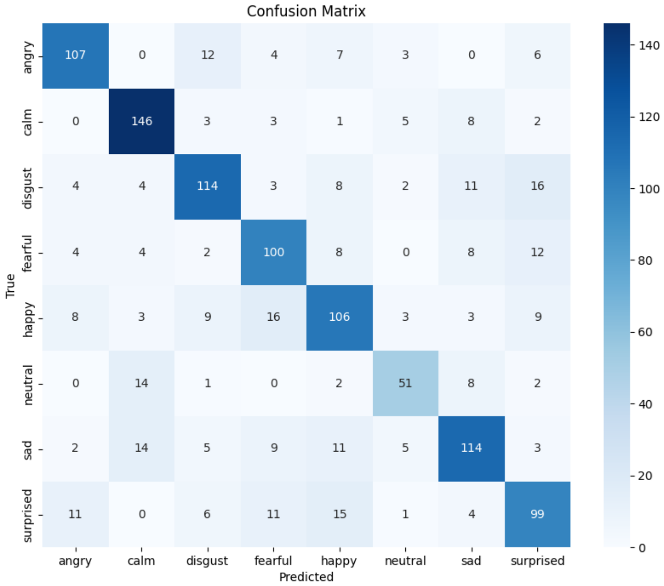
Model Enhancement Attempt 2: Model Architecture and Training Changes - Results





Accuracy of 73%, reached goal!

Model Enhancement Attempt 2: Model Architecture and Training Changes - Results

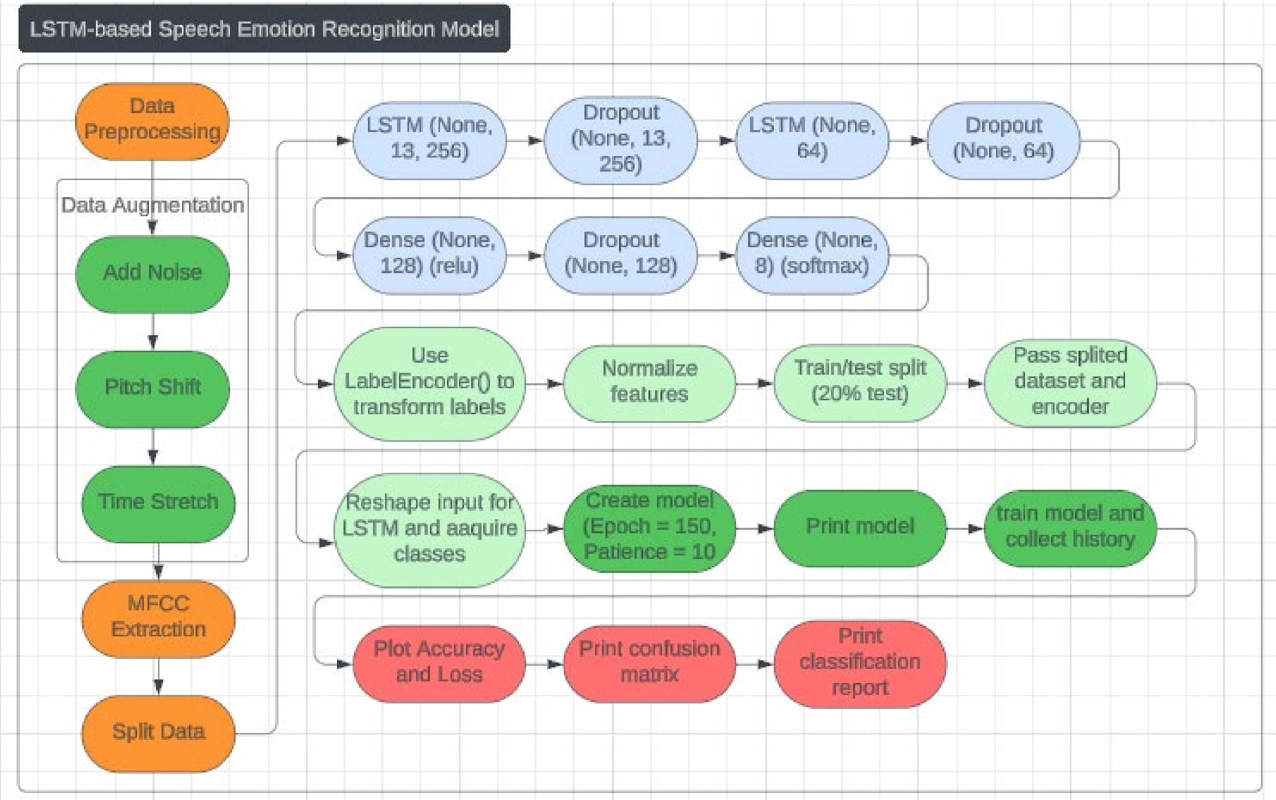



```
Test Loss: 1.0247074365615845
Test Accuracy: 0.7265625
36/36 ━━━━━━━━━━━━━━━━━━━━ 1s 9ms/step
Classification Report:
              precision    recall  f1-score   support

       angry       0.79      0.77      0.78       139
        calm       0.79      0.87      0.83       168
     disgust       0.75      0.70      0.73       162
     fearful       0.68      0.72      0.70       138
       happy       0.67      0.68      0.67       157
     neutral       0.73      0.65      0.69        78
         sad       0.73      0.70      0.71       163
   surprised       0.66      0.67      0.67       147

    accuracy                           0.73      1152
   macro avg       0.73      0.72      0.72      1152
weighted avg       0.73      0.73      0.73      1152
```

Putting it all together: Final Model



Final Though

- It is much easier for us to detect emotions from voice for close people compared with strangers. We learn how a loved one sounds when they are sad, happy, etc.
- THUS, models need to consistently learn! If an SER is used in gaming, each player may have their own variant of the model weight that evolves and adapts with the player, hence converging to a near perfect accuracy over time.

Demo

Summary Table:

	File	Actual Emotion	Predicted Emotion	Result
0	./RAVDESS/Actor_02/03-01-01-01-01-01-02.wav	neutral	happy	Incorrect
1	./RAVDESS/Actor_02/03-01-02-02-01-01-02.wav	calm	happy	Incorrect
2	./RAVDESS/Actor_02/03-01-03-01-02-01-02.wav	happy	happy	Correct
3	./RAVDESS/Actor_03/03-01-04-01-01-01-03.wav	sad	sad	Correct
4	./RAVDESS/Actor_03/03-01-05-02-01-01-03.wav	angry	angry	Correct
5	./RAVDESS/Actor_04/03-01-06-01-02-01-04.wav	fearful	happy	Incorrect
6	./RAVDESS/Actor_04/03-01-07-02-02-01-04.wav	disgust	surprised	Incorrect
7	./RAVDESS/Actor_05/03-01-08-01-01-01-05.wav	surprised	surprised	Correct
8	./RAVDESS/Actor_06/03-01-02-01-02-01-06.wav	calm	calm	Correct
9	./RAVDESS/Actor_06/03-01-03-02-01-01-06.wav	happy	happy	Correct
10	./RAVDESS/Actor_07/03-01-04-01-01-01-07.wav	sad	calm	Incorrect
11	./RAVDESS/Actor_07/03-01-05-01-02-01-07.wav	angry	angry	Correct
12	./RAVDESS/Actor_08/03-01-06-02-01-01-08.wav	fearful	happy	Incorrect
13	./RAVDESS/Actor_08/03-01-07-01-01-01-08.wav	disgust	disgust	Correct
14	./RAVDESS/Actor_09/03-01-08-02-01-01-09.wav	surprised	fearful	Incorrect
15	./RAVDESS/Actor_09/03-01-01-01-02-01-09.wav	neutral	sad	Incorrect
16	./RAVDESS/Actor_10/03-01-02-02-02-01-10.wav	calm	sad	Incorrect
17	./RAVDESS/Actor_10/03-01-03-01-01-01-10.wav	happy	disgust	Incorrect
18	./RAVDESS/Actor_11/03-01-04-02-01-01-11.wav	sad	sad	Correct

Q&A

Thank you!