Multimodal Emotion Recognition from Audio

Arnav Govindu
ME23BTECH11009
me23btech11009@iith.ac.in

Abstract

This project explores multimodal emotion recognition from audio using both spectral and prosodic features. Mel spectrograms are processed via a convolutional neural network (CNN), while pitch, loudness, and tempo are input to a separate feedforward neural network. A late-fusion architecture combining these modalities significantly outperforms the unimodal models on the RAVDESS dataset.

Dataset

The RAVDESS Emotional Speech Audio dataset contains 1,440 speech clips labeled with 8 emotion categories. Each audio sample is a .wav file encoded with metadata including emotion, actor, and modality.

Phase I: Spectrogram-Based Modeling

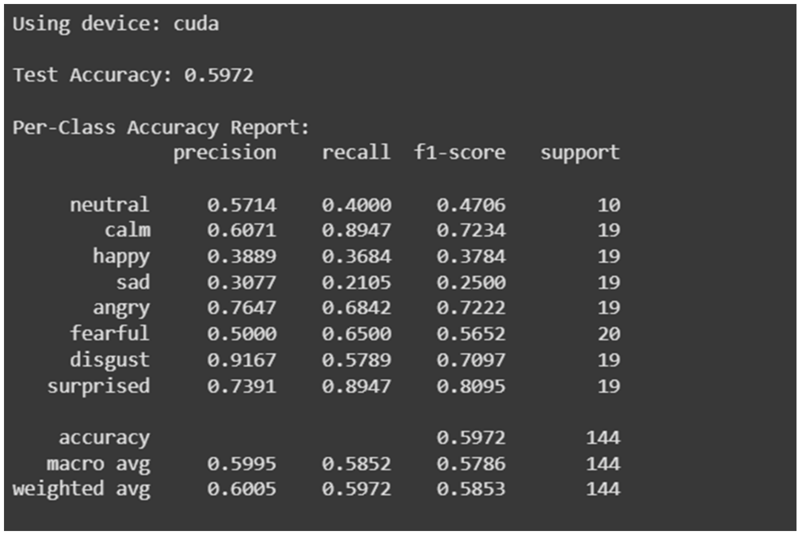
Data Preparation

- Audio files were recursively loaded using full paths.
- Converted to mel spectrograms with Librosa and normalized.
- Spectrograms were padded or truncated to a fixed time length.
- Each sample is a tuple of (spectrogram tensor, emotion label).

CNN Model

A CNN was trained to classify emotions from spectrograms.

- Accuracy: 57% (local GPU), 60% (Google Colab)



```
Using device: cuda

Test Accuracy: 0.5972

Per-Class Accuracy Report:
              precision    recall  f1-score   support

     neutral     0.5714    0.4000    0.4706        10
        calm     0.6071    0.8947    0.7234        19
       happy     0.3889    0.3684    0.3784        19
         sad     0.3077    0.2105    0.2500        19
       angry     0.7647    0.6842    0.7222        19
     fearful     0.5000    0.6500    0.5652        20
     disgust     0.9167    0.5789    0.7097        19
   surprised     0.7391    0.8947    0.8095        19

    accuracy                         0.5972       144
   macro avg     0.5995    0.5852    0.5786       144
weighted avg     0.6005    0.5972    0.5853       144
```

Figure 1: Training accuracy/loss for CNN model

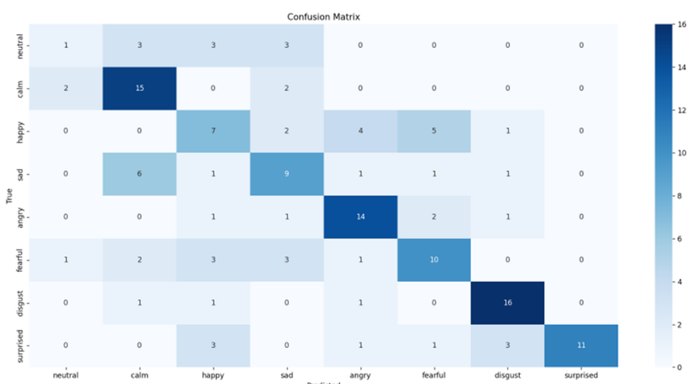


Figure 2: Confusion matrix - CNN model

Evaluation and Inferences

`eval.py` loads the best-performing CNN model checkpoint and computes evaluation metrics. It uses `scikit-learn` for metrics and `matplotlib/seaborn` for visualization.

The CNN performed well on calm, angry, disgust, and surprised emotions. These likely have distinct spectral signatures—e.g., high energy for angry/surprised, and smooth pitch patterns for calm. The model struggled with neutral, happy, sad, and fearful, possibly due to overlapping or subtle spectral cues.

All further models were trained on my local (laptop) GPU

Phase II: Prosodic Feature Modeling

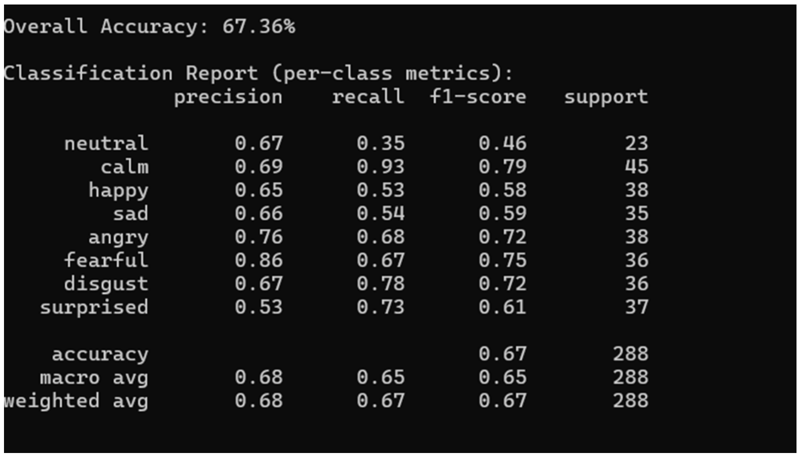
Feature Extraction

Prosodic features were extracted using Librosa:

- Pitch and loudness sequences (padded to fixed length)
- Single scalar tempo value
- Combined and normalized into a 1D vector

Prosody Model (MLP)

- Architecture: Feedforward neural network
- Accuracy: 67%



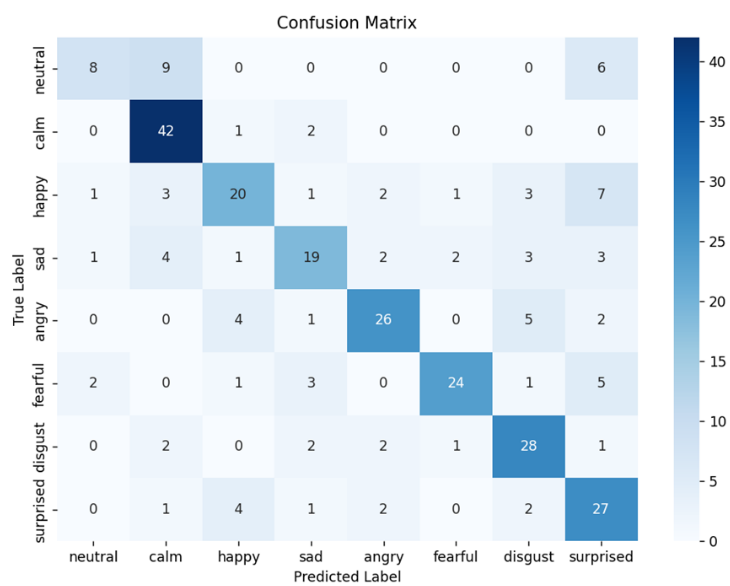
```
Overall Accuracy: 67.36%

Classification Report (per-class metrics):
              precision    recall  f1-score   support

     neutral       0.67      0.35      0.46        23
        calm       0.69      0.93      0.79        45
       happy       0.65      0.53      0.58        38
         sad       0.66      0.54      0.59        35
       angry       0.76      0.68      0.72        38
     fearful       0.86      0.67      0.75        36
     disgust       0.67      0.78      0.72        36
   surprised       0.53      0.73      0.61        37

    accuracy                           0.67       288
   macro avg       0.68      0.65      0.65       288
weighted avg       0.68      0.67      0.67       288
```

Figure 3: Training accuracy/loss - Prosody model



True \ Predicted	neutral	calm	happy	sad	angry	fearful	disgust	surprised
neutral	8	9	0	0	0	0	0	6
calm	0	42	1	2	0	0	0	0
happy	1	3	20	1	2	1	3	7
sad	1	4	1	19	2	2	3	3
angry	0	0	4	1	26	0	5	2
fearful	2	0	1	3	0	24	1	5
disgust	0	2	0	2	2	1	28	1
surprised	0	1	4	1	2	0	2	27

Figure 4: Confusion matrix - Prosody model

Inferences

The model performed well on calm (42 correct), disgust (28), and surprised (27). This aligns with the characteristic low energy and smooth pitch contours often associated with calm affect, which likely translated to consistent patterns in the prosodic features. It struggled with neutral, happy, and sad due to their less distinct or overlapping prosodic cues.

Phase III: Fusion Model

Fusion Architecture

CNN and prosody models were repurposed as feature extractors:

- CNN: Output from final convolutional layer (flattened)
- Prosody model: Output from first fully connected layer

These vectors were concatenated and passed to a new MLP classifier.

Results

- Accuracy: 90.28%

Fusion Model Accuracy: 0.9028

Classification Report:

	precision	recall	f1-score	support
neutral	0.8333	0.9375	0.8824	16
calm	0.8788	0.8056	0.8406	36
happy	0.8056	0.8529	0.8286	34
sad	0.8444	0.9048	0.8736	42
angry	0.9459	0.9459	0.9459	37
fearful	0.8936	0.8750	0.8842	48
disgust	1.0000	0.9524	0.9756	42
surprised	1.0000	0.9697	0.9846	33
accuracy			0.9028	288
macro avg	0.9002	0.9055	0.9019	288
weighted avg	0.9053	0.9028	0.9033	288

Figure 5: Fusion model training accuracy

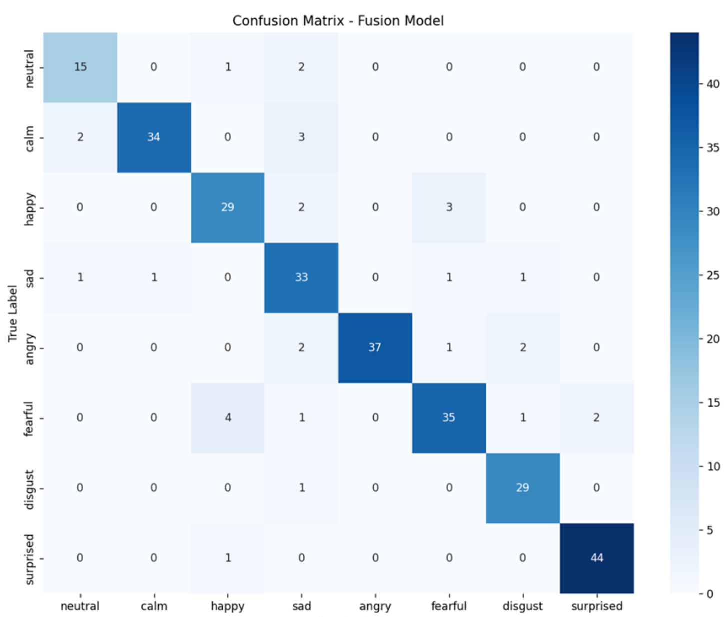


Figure 6: Confusion matrix - Fusion model

Evaluation and Inferences

`evaluate_prosody.py` loads the best fusion model and computes metrics on the validation set.

The fusion model improved performance on all categories, especially happy and neutral. Sad remained challenging, indicating subtle prosodic and spectral cues.

Comparison of Models

Model	Accuracy (%)	Best Class	Worst Class
CNN (Spectrograms)	60.0	Angry	Neutral
MLP (Prosody)	67.0	Calm	Neutral
Fusion Model	90.3	Disgust	Sad

Table 1: Performance comparison of models

Conclusion

The late-fusion approach significantly improved emotion recognition performance over unimodal baselines. This highlights the complementary nature of spectral and prosodic information in emotion classification.

References

[1] Brian McFee et al., *librosa: Audio and Music Signal Analysis in Python*, 2015.

[2] Pedregosa et al., *Scikit-learn: Machine Learning in Python*, JMLR, 2011.

[3] RAVDESS Dataset.