Emotion Recognition Pipeline: Final Task Report

May 08, 2025

Author: Gagan Chandra

1. Introduction

This report presents a robust emotion recognition pipeline developed using the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). The pipeline integrates unimodal and multimodal deep learning models to classify eight emotions: neutral, calm, happy, sad, angry, fearful, disgust, and surprised. The implementation includes:

- **Unimodal Models**: Audio CNN for mel-spectrograms, Text RNN for transcribed text.
- **Multimodal Model**: Combines audio and text features for enhanced performance.

The project leverages PyTorch, torchaudio, and the Whisper-tiny model for transcription, with comprehensive evaluation metrics and visualizations to assess model performance.

2. Pipeline Overview

2.1 Timeline

The pipeline was developed in distinct phases, visualized below using a timeline.

3. Methodology

3.1 Dataset

The RAVDESS dataset comprises 1440 audio files from 24 actors (12 male, 12 female), with each actor contributing 60 files across eight emotions. The dataset is balanced, with approximately 180 samples per emotion.

3.2 Preprocessing

- **Audio**: Converted to mel-spectrograms (128 mel bins, 280 frames) using torchaudio's `MelSpectrogram`. Waveforms were resampled to 48kHz and padded/truncated to 30 seconds.
- **Text**: Transcribed using Whisper-tiny, tokenized into word indices with a vocabulary size of approximately 200 words. Maximum transcript length is 6 words.
- **Split**: 80% training (1152 samples), 20% testing (288 samples), stratified by emotion.

3.3 Models

- **Audio CNN**: Six convolutional layers (32, 64, 128, 256, 512, 512 filters) with batch normalization, followed by three fully connected layers (512→256→128→8). Uses max-pooling and dropout (0.5).
- **Text RNN**: Bidirectional LSTM (2 layers, 128 hidden units) with 64-dimensional embeddings, followed by two fully connected layers (256→128→8).
- **Multimodal**: Concatenates audio (256 features) and text (128 features) into a 384-dimensional vector, processed by three fully connected layers (384→256→128→8) with batch normalization.

3.4 Training

Models were trained for 50 epochs with a batch size of 16, using the Adam optimizer (learning rate 0.001) and CrossEntropyLoss. Gradient accumulation (2 steps) and mixed precision training (via `torch.amp`) were employed to optimize GPU memory usage on an RTX 4090.

3.5 Evaluation

Performance was assessed using accuracy, precision, recall, and F1-score (weighted averages). Confusion matrices and training loss curves were generated to analyze model behavior. All metrics were computed on the test set (288 samples).

4. Results

4.1 Training and Validation Accuracies

Table 1: Test Performance of Models

Model	Test Accuracy (%)	Precision	Recall	F1-Score
Audio CNN	69.28	0.69	0.68	0.68
Text RNN	13.12	0.13	0.12	0.13
Multimodal	60.14	0.61	0.60	0.60

5. Inferences

5.1 Audio CNN

The Audio CNN achieved a test accuracy of 69.28%, excelling in emotions with distinct acoustic patterns, such as angryand happy . However, it struggled with neutral and calm emotions due to their subtle acoustic differences, as evidenced by lower F1-scores.

5.2 Text RNN

The Text RNN's low accuracy (13.12%) reflects the limitations of short transcripts (maximum 6 words), which provide insufficient semantic context. Emotions like angry and surprised, with more distinctive phrasing, had slightly better F1-scores, but overall performance was hindered by the small vocabulary and transcript brevity. The confusion matrix indicates near-random predictions for neutral and calm.

5.3 Multimodal Model

The Multimodal model outperformed unimodal models with a test accuracy of 60.14%, leveraging complementary audio and text features. It showed significant improvements in angry and happy , where acoustic and textual cues align. Subtle emotions like neutral remained challenging, but the model's robustness was evident in reduced misclassifications

5.4 Challenges

- **Dataset Duplication**: Initial file counts exceeded 1440 due to duplicate folders. Deduplication logic prioritized the /usr/Downloads directory, but manual verification is recommended.
- **Transcription Errors**: Whisper-tiny struggled with short, emotionally charged audio, leading to incomplete transcripts for some samples.
- **Model Complexity**: Deeper architectures (e.g., 6-layer CNN) risked overfitting, mitigated by dropout (0.5) and batch normalization.
- **Shape Mismatch**: A bug in the Multimodal model (384 vs. 768 features) was fixed by correcting the feature concatenation (256 audio + 128 text).

6. Conclusion

The emotion recognition pipeline effectively classifies eight emotions using unimodal and multimodal approaches. The Multimodal model's superior performance (60.14% accuracy) highlights the benefit of combining audio and text modalities. Key findings include:

- Audio features are critical for emotions with strong acoustic signatures (e.g., angry, happy).
- Text features are limited by short transcripts but enhance performance when combined with audio.
- Subtle emotions (neutral, calm) remain challenging, requiring further feature engineering.

Future improvements include:

- Fine-tuning Whisper for better transcription accuracy.
- Implementing data augmentation (e.g., pitch shifting) to enhance robustness.
- Using a bigger neural network architecture.
- Incorporating Transformers.

Thank you !