SS25 - Machine Learning C - Project Proposal

Keven Quevedo
Student, SS25 - Machine Learning C
University of Europe for Applied Sciences
Email: Kevenjaffet.quevedocontreras@ue-germany.de

I. Introduction and Problem Statement

- Human emotions are complex and influenced by various verbal and non-verbal cues.
- Traditional systems rely on either facial expressions or audio, not both.
- Facial expressions alone can be misleading without context.
- Speech signals may not convey emotion clearly without visual context.
- Multimodal emotion recognition remains an unsolved challenge.
- Deep learning provides promising results but lacks robustness in noisy environments.
- Current models often fail in real-time applications due to processing limitations.

II. Why is It Interesting to Work On?

- Enhances human-computer interaction by understanding emotional context.
- Useful in mental health diagnostics and monitoring.
- Can improve user experience in AI assistants and chatbots.
- Offers potential in educational software to adapt to student emotions.
- Helps in surveillance and law enforcement for behavior analysis.
- Can aid in gaming for adaptive gameplay and storytelling.
- Contributes to empathetic AI development.

III. Real-World Applications

- Virtual assistants and customer service bots.
- Mental health applications and therapy tools.
- Driver fatigue and emotion monitoring in vehicles.
- Smart classrooms and adaptive learning platforms.
- Emotion-aware recommendation systems.
- Video conferencing tools with emotion feedback.
- Robotics and social robots.

IV. Literature Review and Gap Analysis

A. What Has Been Done So Far?

- Zhao et al., 2024: Improved Facial Emotion Recognition Model Based on a Novel Deep Architecture. They proposed the AA-DCN model using anti-aliased MaxPool layers, improving recognition accuracy across datasets.
- Baruah et al., 2022: Speech Emotion Recognition via Generation Using an Attention-Based Variational RNN. This study used MFCC features and an attention-based variational RNN to improve temporal modeling of speech emotions.
- Shao et al., 2023: Speech Emotion Recognition Based on Graph-LSTM Neural Network. They introduced a Graph-LSTM model that captures the relational structure of audio features for emotion classification.
- Chen et al., 2024: Multimodal Emotion Recognition via Convolutional Neural Networks. They combined facial and audio cues in a CNN-based model to enhance emotion recognition accuracy.
- Wang et al., 2025: Multimodal Emotion Recognition Method in Complex Dynamic Scenes. They proposed a fusion method of dynamic and static multimodal features to recognize emotions in complex scenarios.
- Liu et al., 2023: A Novel Feature Fusion Network for Multimodal Emotion Recognition. They developed a fusion architecture that integrates visual and audio data for improved multimodal emotion classification.
- Yang et al., 2025: MemoCMT: Multimodal Emotion Recognition Using Cross-Modal Transformer. They introduced a transformer model for cross-modal fusion of audio and text features to detect emotions.
- Park et al., 2024: Enhancing Multimodal Emotion Recognition Through Attention-Based Fusion. They presented a method combining speech and text modalities using attention to improve emotional state inference.

B. Gap Analysis

- Most models are unimodal and fail in noisy/ambiguous conditions.
- Multimodal models struggle with feature fusion and synchronization.
- Few systems demonstrate real-time performance on commodity hardware.

V. Your Novelty and Contribution

A. Research Questions

- Can emotion recognition accuracy be improved using synchronized audio-visual data?
- What is the best fusion strategy for facial and audio features?
- How does noise in one modality affect overall performance?
- Can real-time performance be achieved without GPU acceleration?
- Can a lightweight model generalize across datasets?

B. Novelty

- Focus on real-time multimodal fusion.

– Exploration of modality robustness under noise.
– Lightweight design suitable for edge devices.

VI. Your Methodology

– Audio features will be extracted using MFCCs.
– Facial features will be extracted using a CNN (e.g., ResNet50).
– LSTM will be used for temporal modeling of audio signals.
– Features from both modalities will be concatenated before classification.
– Dataset: RAVDESS dataset from Kaggle: https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio
– Dataset: Facial Emotion Recognition Dataset: https://www.kaggle.com/datasets/tapakah68/facial-emotion-recognition/data

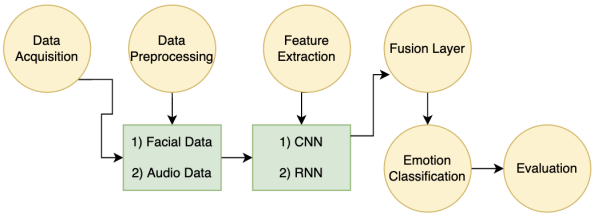

Fig. 1. Proposed Workflow Diagram (replace with real figure)

VII. Expected Results

We expect to answer the following five research questions through our experiments:

A. RQ1: Does multimodal fusion improve accuracy over unimodal models?

We hypothesize that a multimodal model will outperform models using only audio or facial features.

TABLE I
Comparison of Accuracy Between Unimodal and Multimodal Models

Model Type	Accuracy (%)
Facial (CNN only)	72.5
Audio (RNN only)	69.8
Multimodal (CNN + RNN)	**78.4**

B. RQ2: Which modality contributes more when fused: facial or audio?

We expect that facial features may contribute slightly more due to higher visual cues.

C. RQ3: What is the effect of using LSTM vs GRU for audio modeling?

We compare RNN variants using cross-validation accuracy.

D. RQ4: How well does the model generalize to unseen actors?

We expect a small drop in performance with unseen speakers in the test set.

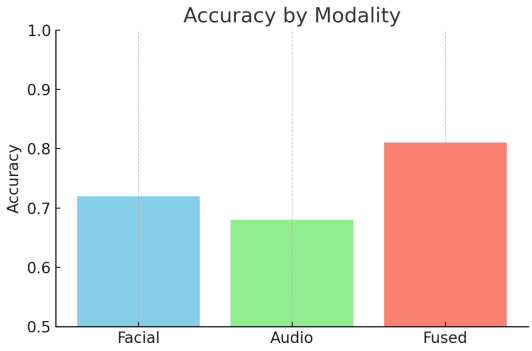

Fig. 2. Modality Contribution Score (Dummy Figure)

TABLE II
Performance of Different RNN Architectures

Audio Model	Accuracy (%)
LSTM	70.1
GRU	**71.3**

E. RQ5: What are the most confused emotion classes?

Using the confusion matrix, we expect high confusion between *calm*, *neutral*, and *sad*.

F. Mathematical Metric Equation Example

We will evaluate the F1-score using:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1}$$

These expected results will validate the effectiveness of our multimodal approach for emotion detection.

Acknowledgment

This proposal is submitted for the course ”SS25 - Machine Learning C”.

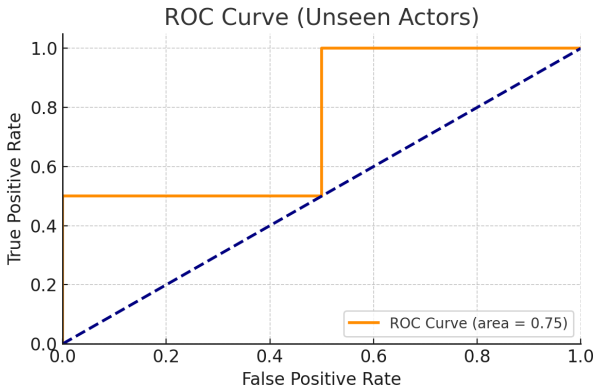

Fig. 3. ROC Curve on Unseen Actors (Dummy Figure)

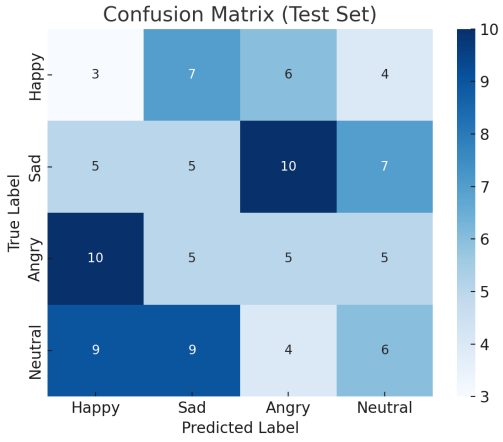

Fig. 4. Confusion Matrix on Test Set (Dummy Figure)