Emotion-Aware AI Companion

***Micah Baldonado**
Biomedical Engineering
mbaldona@andrew.cmu.edu

Aryaman Shandilya
Civil and Environmental Engineering
aryamans@andrew.cmu.edu

Peter Ragone
Electrical and Computer Engineering
pragone@andrew.cmu.edu

Prajwal Kumar
Information Networking Institute
prajwalk@andrew.cmu.edu

1 Introduction

We aim to create a multimodal empathetic AI therapist and emotional companion that synthesizes text sentiment, facial emotion sentiment, and speech emotion sentiment. While text and facial sentiment components leverage publicly available pretrained models, our speech emotion recognition (SER) model is developed from scratch as the primary deep learning contribution for this project. The long-term goal is to integrate all three modalities into a unified architecture where an LLM-based agent dynamically adapts its responses based on the user's emotional state. As of midterm, we have independently implemented: (1) a working emotional dialogue system powered by an LLM, (2) facial emotion detection using DeepFace's pretrained CNN-based model, and (3) a speech emotion classifier with promising baseline performance. The components are functional in isolation, and integration is ongoing.

2 Related Work

Facial emotion recognition using CNNs has shown robust performance in healthcare and teleconsultation settings (1), with models like FacialNet (5) trained on datasets such as FER2013 (4) offering reliable seven-class emotion classification. For text-based dialogue, pretrained language models have been adapted to reflect emotional nuance by grounding responses in curated sources such as psychotherapy texts (2) and the EmpatheticDialogues dataset (3).

In speech emotion recognition (SER), traditional pipelines use features like MFCCs and mel spectrograms to train models such as multilayer perceptrons (MLPs) or convolutional neural networks (CNNs). We follow a similar structure, training a CNN on mel spectrograms extracted from the RAVDESS dataset. Our baseline model achieves 49% accuracy and provides a strong foundation for further improvement. This aligns with recent SER research such as the method proposed by Kim and Lee (6), which utilizes vision transformers and positional encoding to model temporal-frequency correlations in spectrograms for enhanced SER performance.

3 Methodology

Our system is built on three core sentiment recognition modules: text, facial, and speech. Each feeds into a centralized LLM agent that includes two internal components—an analyzer and a responder—to produce emotionally attuned output.

Text Sentiment: We use a pretrained transformer model fine-tuned on emotional dialogue datasets. The model outputs one of several predefined emotion classes per user message.

Facial Emotion Sentiment: We use DeepFace, a publicly available pretrained emotion recognition framework, to extract facial emotion in real-time from webcam input. The system captures frames using OpenCV and passes them to DeepFace, which outputs the dominant facial emotion based on its internal ensemble of CNNs trained on datasets such as FER2013, AffectNet, and others.

Speech Emotion Sentiment: We trained a CNN-based classifier on mel spectrograms of audio clips.

- **Dataset**: CREMA-D was used as the dataset of choice. It contains 7,442 unique audio clips from 91 actors with varying accents. The sentences were presented using one of six different emotions (Anger, Disgust, Fear, Happy, Neutral and Sad)(7).
- **Model Architecture (Current)**: We are using Mel spectograms of audio files as inputs and passing them through two convolutional layers with ReLU activations and max pooling, followed by a fully connected classification head. MFCCs were explored as the input in earlier iterations but Mel Spectrograms offered richer input representations.
- **Hyperparameters**: Training was conducted over 10 epochs using Adam optimizer and CrossEntropyLoss on an Apple Silicon GPU.
- **Future Plan**: This baseline will be improved through augmentation (e.g., noise injection, pitch shift), hyperparameter tuning, and eventually transfer learning using models like Wav2Vec 2.0.

LLM Agent: Our central language model is a working implementation of an emotional companion system. It receives multimodal emotion labels and user input to guide response generation. The analyzer processes emotional signals, and the responder produces empathetic dialogue. Our goal is not traditional accuracy metrics, but maximizing emotional coherence and perceived empathy.

4 Results

The core focus of our results is on speech emotion recognition as the primary deep learning component of this project. During our experiments, we achieved the following results:

Model	Feature Type	Test Accuracy	Notes
MLP	MFCC	39.8%	Baseline
CNN	Mel Spectrogram	49.0%	Expected to perform better due to richer input

5 Ablation Study (if applicable)

Our core goal is to build a cohesive multimodal emotional companion. While ablation is not our primary focus, we plan to evaluate how removal of each modality affects output quality. This includes testing the LLM agent's response coherence and empathy when relying solely on text, facial, or speech sentiment if time permits.

6 Discussion and Conclusion

This project emphasizes both a strong deep learning implementation and a compelling emotional AI product. Speech emotion recognition has been our most technically challenging component, but we have achieved a 49% baseline accuracy using a CNN trained on mel spectrograms. We plan to improve performance with advanced architectures and augmentation. At the same time, our LLM agent shows promise as an emotionally responsive dialogue system. Integrating the three modalities will help us approach a system that not only understands affect but adapts its behavior accordingly. Challenges remain in real-time fusion, dataset variability, and measuring perceived empathy, but our progress demonstrates feasibility and sets the foundation for a novel therapeutic AI system.

References

[1] Hayette Hadjar, Binh Vu, and Matthias Hemmje. Therasense: Deep learning for facial emotion analysis in mental health teleconsultation. *Electronics*, 14(3), 2025.

[2] International Psychotherapy Institute. Free ebooks in psychotherapy, psychiatry and psychoanalysis, 2025. Accessed: March 15, 2025.

[3] Athar Jairath. Empathetic dialogues (facebook ai) 25k, 2022. Accessed: March 16, 2025.

[4] Yousif Khaireddin and Zhuofa Chen. Facial emotion recognition: State of the art performance on fer2013. *arXiv preprint*, cs.CV, 2021.

[5] In-seop Na, Asma Aldrees, Abeer Hakeem, Linda Mohaisen, Muhammad Umer, Dina Abdulaziz Al-Hammadi, Shtwai Alsubai, Nisreen Innab, and Imran Ashraf. Facialnet: Facial emotion recognition for mental health analysis using user segmentation with transfer learning model. *Frontiers in Computational Neuroscience*, 18, 2024.

[6] Jeong-Yoon Kim and Seung-Ho Lee. Accuracy Enhancement Method for Speech Emotion Recognition from Spectrogram Using Temporal Frequency Correlation and Positional Information Learning Through Knowledge Transfer. *arXiv preprint arXiv:2403.17327*, 2024.

[7] Cao H, Cooper DG, Keutmann MK, Gur RC, Nenkova A, Verma R. CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset. *IEEE Trans Affect Comput. 2014 Oct-Dec;5(4):377-390. doi: 10.1109/TAFFC.2014.2336244. PMID: 25653738; PMCID: PMC4313618.*

[0] *Micah is working on a shared component of this project between this class and 42698.