Multimodal Mood Recognition (Text + Voice) for AI Companions

Israr Qayyum
Email: israrqayyumf22@nutech.edu.pk

Mehdi Ali
Email: mehdialif22@nutech.edu.pk

I. Introduction

The evolution of Human-Computer Interaction (HCI) has shifted from command-based interfaces to conversational agents designed to function as intelligent companions. A critical requirement for these systems is the ability to perceive and respond to human emotions, a domain known as Affective Computing. While early systems relied primarily on textual sentiment analysis, human communication is inherently multimodal; emotions are conveyed not only through lexical choices but also through prosodic cues such as pitch, tone, and speaking rate. Consequently, relying solely on text often leads to misinterpretation—for example, the phrase "I am fine" may indicate contentment or distress depending entirely on the vocal delivery. To address this, the specific technical domain of Multimodal Emotion Recognition (MER) has emerged, focusing on the fusion of linguistic and acoustic data to create more empathetic and responsive AI systems.

Recent academic efforts have sought to bridge the gap between text and speech processing. Significant research has focused on fusing representations from large pretrained models. For instance, Zhao *et al.* (2022) utilized transfer learning by combining Wav2Vec 2.0 and BERT, experimenting with multi-granularity fusion (frame, phone, and word levels) to capture fine-grained alignment [1]. While effective on acted datasets, their approach highlights the difficulty of aligning modalities in spontaneous speech. In contrast, Ferreira *et al.* (2025) proposed a complex graph-based ensemble utilizing multiple self-supervised audio encoders (including HuBERT) alongside text models [2]. Although this method improves robustness, the high computational complexity renders it less suitable for real-time applications. Similarly, Deng *et al.* (2025) introduced Sync-TVA, a framework that constructs heterogeneous graphs across text, audio, and visual nodes [3]. While achieving state-of-the-art results, these systems often rely heavily on visual data (facial expressions), which is not always available or privacy-compliant in conversational AI companion settings.

Despite these advancements, a critical gap remains in the development of lightweight, robust mood recognition systems for purely conversational (text and audio) contexts. Existing high-performing models either rely on computationally expensive ensembles, require auxiliary visual modalities (video/face) that are unavailable in many chat interfaces, or lack effective adaptive weighting mechanisms to handle the noise and asynchrony inherent in spontaneous human dialogue. There is a lack of architectures that efficiently align and fuse text and voice cues dynamically while maintaining the low latency required for real-time AI companionship.

A. Problem Statement

The specific problem addressed in this study is the inability of current unimodal or rigid multimodal systems to accurately classify user mood in real-time conversational settings. When users express complex emotions where text and tone contradict, existing AI companions often fail to detect the underlying mood due to the lack of adaptive cross-modal alignment. This study addresses the technical challenge of jointly analyzing textual messages and vocal expressions to produce a unified, accurate mood classification (e.g., happy, sad, angry, neutral) without relying on visual input or computationally prohibitive architectures.

B. Research Contributions

In this study, the key contributions are as follows:

- **Contribution 1:** We propose a streamlined multimodal deep learning framework that effectively integrates pretrained linguistic encoders (BERT) and acoustic encoders (Wav2Vec 2.0) to perform robust mood detection using only text and voice.
- **Contribution 2:** We design and implement an adaptive cross-modal attention mechanism that dynamically weights the importance of text versus audio features, ensuring accurate classification even when one modality is noisy or ambiguous.
- **Contribution 3:** We provide an empirical evaluation of the proposed model on naturalistic conversational datasets, demonstrating improved generalization and accuracy compared to unimodal baselines and standard concatenation fusion methods.

II. Literature Review

Multimodal emotion recognition is a rapidly expanding field, with methodologies generally categorized by their fusion strategies: early fusion, late fusion, and hybrid attention-based fusion.

A. Fusion Strategies and Architectures

Early fusion approaches attempt to concatenate features from different modalities at the input level before feeding them into a classifier. While conceptually simple, this method often struggles with the ”curse of dimensionality” and the synchronization of heterogeneous feature rates (e.g., aligning audio frames with text tokens). To mitigate this, Zhao *et al.* [1] introduced a multi-level fusion strategy using co-attention mechanisms between Wav2Vec 2.0 and BERT. Their work demonstrated that aligning features at the phone and word level yields higher accuracy than simple sentence-level concatenation. However, their reliance on clean, acted data (IEMOCAP) limits the transferability of their findings to the noise inherent in real-world AI companion interactions.

Late fusion, conversely, processes each modality independently and combines the final decision scores. Ferreira *et al.* [2] employed this strategy via a graph-based ensemble. By training separate encoders for prosodic features, spectral features, and text, they achieved robustness against environmental noise. However, late fusion often fails to capture the subtle interplay between “what is said” and “how it is said” (e.g., sarcasm), which requires interaction at the feature representation level.

B. Graph and Attention-Based Frameworks

Recent state-of-the-art systems have moved toward graph neural networks (GNNs) and attention mechanisms to model cross-modal dependencies explicitly. Deng *et al.* [3] proposed Sync-TVA, a graph-attention framework that treats text, audio, and video segments as nodes in a hypergraph. This allows the model to learn dynamic relationships between modalities. While highly effective, the computational cost of building and traversing these graphs is significant. Furthermore, Farhadipour *et al.* [5] highlighted that while adding visual modalities (video) maximizes accuracy, it introduces privacy concerns and hardware dependencies that are often impractical for lightweight mobile AI companions.

C. Summary of Gaps

The literature indicates a clear trade-off: simple models lack nuance, while complex graph-based models lack efficiency. Most importantly, few studies focus specifically on the *adaptive* weighting of text and audio for conversational agents where visual data is absent. This study aims to fill this specific void by developing an efficient, attention-based fusion model tailored for the constraints of AI companionship.

References

[1] Z. Zhao, Y. Wang, and Y. Wang, “Multi-level Fusion of Wav2vec 2.0 and BERT for Multimodal Emotion Recognition,” in *Proc. Interspeech 2022*, Incheon, Korea, 2022, pp. 4725–4729. [Online]. Available: https://doi.org/10.21437/Interspeech.2022-10230

[2] A. I. S. Ferreira, L. R. S. Gris, L. Oliveira, D. Ribeiro, L. Fernando, and F. Lustosa, “Enhancing Speech Emotion Recognition with Graph-Based Multimodal Fusion and Prosodic Features for the Speech Emotion Recognition in Naturalistic Conditions Challenge,” in *Proc. Interspeech 2025*, 2025. [Online]. Available: https://arxiv.org/abs/2506.02088

[3] Z. Deng, Y. Lu, J. Liao, S. Wu, and C. Wei, “Sync-TVA: A Graph-Attention Framework for Multimodal Emotion Recognition with Cross-Modal Fusion,” *arXiv preprint arXiv:2507.21395*, 2025. [Online]. Available: https://arxiv.org/abs/2507.21395

[4] C. Wu, Y. Cai, Y. Liu, P. Zhu, Y. Xue, Z. Gong, J. Hirschberg, and B. Ma, “Multimodal Emotion Recognition in Conversations: A Survey of Methods, Trends, Challenges and Prospects,” in *Findings of the Association for Computational Linguistics: EMNLP 2025*, Suzhou, China, 2025, pp. 6257–6274. [Online]. Available: https://aclanthology.org/2025.findings-emnlp.332/

[5] A. Farhadipour, H. Ranjbar, M. Chapariniya, T. Vukovic, S. Ebling, and V. Dellwo, “Multimodal Emotion Recognition and Sentiment Analysis in Multi-Party Conversation Contexts,” *arXiv preprint arXiv:2503.06805*, 2025. [Online]. Available: https://arxiv.org/abs/2503.06805