A Confidence-Weighted Memory-Augmented Multimodal Voice Agent for Real-Time Emotion-Aware Interaction

Vijaya Bharathi A.* and Prashant Nitnaware
Department of Computer Engineering, Pillai College of Engineering, Panvel, India
Email: jagan21it@student.mes.ac.in (V.B.A.), pnitnaware@mes.ac.in (P.N.)

Manuscript received March 11, 2026; revised April 17, 2026; accepted May 2, 2026
*Corresponding author

***Abstract*—Conversational emotion-aware systems are critical in the development of human-centric Artificial Intelligence (AI). Single-modality systems often fail to identify acoustic patterns and text meaning associated with emotional expression. The study proposes an emotion-aware conversational framework that integrates offline-trained speech emotion recognition with a multimodal chatbot pipeline. In the offline phase, three self-supervised speech models, Wav2Vec2 (waveform to vectors version 2), Hidden-Unit Bidirectional Encoder Representations from Transformers (HuBERT), and Whisper, were fine-tuned on the Ryerson Audio-visual Database of Emotional Speech and Song (RAVDESS) emotional speech dataset for multi-class emotion classification. Among them, the fine-tuned Whisper model achieved superior performance and was selected as the primary acoustic emotion encoder. The trained Whisper model was then deployed within a real-time chatbot architecture, where speech input is transcribed using automatic speech recognition and combined with acoustic emotion predictions. A confidence-weighted multimodal fusion strategy integrates audio and text-based emotion cues, followed by an Adaptive Confidence-Weighted Temporal Memory (ACWTM) module to model short-term emotional continuity. The model attained 87.3 percent accuracy and outperforms the unimodal baselines. However, previous emotion-aware models have failed to address the complex issue of cross-modal dependencies and temporal information in an accurate manner due to real-time considerations, resulting in poor recognition results. The current study offers a novelty with new modules, Adaptive Confidence-Weighted Decision fusion Module (ACWDFM) and ACWTM, which play key roles in improving emotion recognition. It achieves enhanced robustness and conversational coherence with stability, offering a scalable framework for real-time empathetic human-AI interaction.**

***Index Terms*—artificial intelligence, conversational system, emotion-aware interaction, memory-augmented voice agent, multimodal voice agent**

I. Introduction

Empathetic conversational agents represent a significant step forward in the quest for human-computer interaction that simulates human modalities [1]. Earlier conversational systems, such as chatbot ELIZA (first chatbot), PARRY (chatbot to simulate the responses of an individual facing paranoia), and ALICE (artificial linguistic internet computer entity), were mainly based on rule-driven pattern matching and hand-engineered features [2]. While these systems marked the beginning of the journey towards modeling human conversations, they failed to provide any real emotional understanding. More complex modular designs, such as XiaoIce, enabled longer and more engaging conversations but relied on extremely complex pipelines and large amounts of annotated data [3]. Despite the overall progress in natural language processing, many existing virtual assistants still focus on semantic intent analysis, with relatively low awareness of the users' emotional states.

Human communication, on the other hand, is a fundamentally multimodal process. Emotional messages are not only encoded in the linguistic content but also in prosody, tone, pitch, rhythm, and intensity. The relatively new area of affective computing, which has been widely covered in recent studies, claims that high-quality emotion recognition requires the fusion of both acoustic and linguistic information [4]. Recent transformer-based speech models [5], such as wav2vec 2.0 and , Hidden-Unit Bidirectional Encoder Representations from Transformers (HuBERT), have shown excellent potential in learning prosodic representations from raw audio signals.

Another study indicates that the deep semantic contextual representations are obtained from the textual transcripts by the pretrained language models, such as Robustly Optimized Bidirectional encoder Representations from Transformers Approach (RoBERTa) [6]. Although multimodal fusion strategies may improve the classification results, most of the existing systems are based on a single utterance and limits with incapability to provide emotional continuity from one dialogue turn to the next. In addition, most of the existing studies assume ideal input conditions and uncertainty or reliability are unattended in real-time scenarios.

Most existing models are inadequate in temporal emotional modeling. In real-world conversations, emotions change dynamically based on the influence of the previous context. Systems that process each utterance independently often result in sudden or inconsistent emotional expressions, which can be detrimental to perceived empathy and trust [7]. In addition, simple fusion techniques may perform poorly when one modality is unreliable, such as in the case of speech recognition errors

or uncertain vocal expressions. It highlights the necessity of a unified and robust framework for context-aware multimodal conversational systems [8].

To tackle these issues, the proposed method presents a memory-augmented system with a multimodal voice agent. It combines transformer-based speech emotion recognition, automatic speech transcription, and text-based emotion classification. The method also uses confidence-driven late fusion and a memory-based cross-turn emotional system. The proposed system derives acoustic and semantic emotion features from speech and transcribed speech. An Adaptive Confidence-Weighted Decision fusion Module (ACWDFM) is used to weigh the relative importance of the two modalities to improve robustness in noisy or uncertain scenarios. Moreover, an Adaptive Confidence-Weighted Temporal Memory (ACWTM) is employed to accumulate the latest predictions to identify the prevailing emotional path for temporal smoothing and stability.

The need to develop an emotion-aware conversational system is critical in modern systems that incorporate multi-modal processing as well as context-maintenance during interaction. Because current approaches are unstable and are unable to maintain emotions, resulting in inconsistent behavior. The proposed approach fulfills these necessities by integrating multi-modal fusion along with context tracking using memory. Thus, it enables empathetic and more precise real-time interaction, maintaining contextual information.

The research goal is to develop and assess a scalable real-time system that connects utterance-level emotion recognition with dialogue-level emotional continuity. Differing from previous methods that focus on isolated classification performance, the proposed system aims to provide better contextual consistency, robustness, and practicality. The scope of the study includes offline benchmarking and an ablation study with user interaction testing.

The significant contributions of the work are:

- A unified multimodal transformer architecture for joint speech and text emotion recognition.
- ACWDFM approach to improve robustness in modality uncertainty.
- ACWTM − A cross-turn emotional memory approach to capture the dynamic trajectories of affect.
- A prototype of a real-time voice agent system that is experimentally validated.

This work contributes to affect-aware conversational AI (artificial intelligence) by going beyond the static utterance-level modeling in the recent state of the art. It relates to the recent research trends in the high-impact scholarly literature.

The structure of the paper contains a literature survey to discuss existing approaches as well as the identified research gap. Further, the proposed methodology covers the novel multimodal approach along with the experiment description and evaluation criteria. The results and discussion covers the performance of the system with offline experiments, and classification, including real-time inference. Comparative analysis discussion precedes the conclusion that discusses limitations and future scope.

II. Literature Survey

Sentiment and affect analysis have been considered a basic need for the development of empathetic and socially intelligent conversational agents. Existing research shows emotion-aware dialogue systems improve user engagement and trust, while perceived intelligence is also enhanced as compared to emotion-neutral dialogue systems. The early approaches using lexicon-based and machine learning mainly relied on text-based sentiment analysis [9]. Although they are more efficient, they overlook critical paralinguistic information such as pitch, tone, intensity, and rate, necessary for effective emotion recognition in speech-based interactions.

A. Related Works

Namratha *et al.* [10] developed an emotion-assisted conversational AI system that incorporates speech recognition, audio-based speech emotion classification, and adaptive response generation. The system uses Wav2Vec2 (waveform to vectors version 2) for speech recognition and a SpeechBrain-based model for speech emotion classification. The system reported class-wise accuracies for neutral, disgust, and anger. The average word error rate confirmed the reliable performance of the transcription module. The system employed detected emotions to inform response generation using large language models, and the expressive responses were synthesized using neural text-to-speech. Although highly effective, the proposed approach did not model cross-turn emotional continuity.

Fulzele *et al.* [11] introduce an emotion-sensing voice assistant that focuses on text-based emotion recognition and adaptive response plans. The assistant identifies emotions, namely happiness, sadness, anger, and surprise, using keyword-based analysis and machine learning algorithms for text analysis. Large language models are used for intent analysis and response formulation. Though adaptable over time, the model's sensitivity to lexical features affects robustness when audio prosodic features contain more intense emotional information than the text. Transformer-based language models have also been explored for sentiment and emotion recognition tasks. Experiments with GPT-2 (generative pre-trained transformer) and RoBERTa [11] suggest that GPT-2 can achieve training accuracy of up to 94%, while validation accuracy can decrease to around 60% on some datasets and 86% on others. RoBERTa generalizes less well in some settings, with training accuracy of about 32-33% and validation accuracy of 6.7% in later epochs. These findings demonstrate the model's sensitivity to dataset properties and the challenges of using text-only models alone.

Rathnayake *et al.* [12] designed an adaptive voice communication system by integrating real-time emotion recognition with GPT-3.5-powered dialogue generation and voice communication. The system connects emotion-sensing text-to-speech with corresponding 3D (3-dimensional) avatar animation to enhance realism and user

engagement. The system is very effective in terms of expressive output and cultural adaptability, but it focuses more on how natural the output responses feel rather than providing a comprehensive multimodal fusion or an emotion model over time.

Hassani and Kangavari [13] conducted a comprehensive survey on emotion-sensing speech generation systems, categorizing existing prosody modeling techniques and identifying challenges such as emotion ambiguity, data scarcity, and inconsistency in evaluation. The survey provides valuable insights into expressive speech synthesis, but does not provide a comprehensive conversational system architecture.

Abdelaziz *et al.* [14] developed an emotion-sensing chatbot system by integrating sentiment analysis with facial expression rendering and real-time lip synchronization. The system achieved an accuracy rate of over 85% in sentiment classification and maintained a lip-sync latency of 50 ms, significantly improving realism and user-perceived empathy. However, the system emphasized visual expressiveness rather than cross-turn emotional reasoning.

Huan Wang and Xiaohui Wang [15] explored multimodal automatic speech recognition with a special emphasis on gender differences using CREMA-D (crowd-sourced emotional multimodal actors dataset) as well as Ryerson audio-Visual Database of Emotional Speech and Song Datasets (RAVDESS) . They experimented with models such as Whisper and achieved high transcription accuracy and improved emotion recognition, especially for females. The research highlights the importance of personalization and demographics, but it does not involve memory-based affect modeling.

Bommireddy Neha *et al.* [16] introduced Aurora, a multi-personality AI voice assistant that incorporates Whisper-based speech recognition, emotive speech synthesis, toxicity detection, and session-level memory. Preliminary findings indicated increased engagement and trust. However, the emotional memory is still bound to individual sessions and does not necessarily smooth out emotional paths across turns in a conversation.

In the research, Khan *et al.* [17] proposed a deep feature fusion-based Multimodal Speech Emotion Recognition (MSER) system utilizing a multi-headed cross-attention approach. The system processed the raw speech with text data independently using Convolutional Neural Networks (CNN) based encoders to obtain distinctive acoustic and semantic features, which were then combined using cross-attention. It improved the interaction between the models while a region-wise deep fusion approach integrated multi-layer features from the audio and text modalities. When tested on the interactive Emotional Dyadic Motion Capture (IEMOCAP) and Multimodal Emotion Lines Dataset (MELD) datasets, the system achieved an approximate 4.5% enhancement over the previous tactics. Pai *et al.* [18] developed a holistic multimodal framework that combines vision using vision transformers and Convolutional Neural Networks (CNNs), text using transformer models like Bidirectional Encoder Representations from Transformers (BERT) and GPT-4, and speech using Recurrent Neural Networks (RNNs) and self-supervised learning methods. The framework used an attention-based fusion strategy to properly align the cross-modal features. The model demonstrated an accuracy improvement by 9% to 12%, a 15% relative improvement in performance on sarcasm recognition, while a relative reduction of 23% in factual error than multimodal CLIP (contrastive language-image pre-training) and Flamingo models.

Previous works have focused on real-time recognition of emotions through multiple modalities, albeit with certain drawbacks. BERT and CNN/transformer encoders were used by Handa *et al.* [19], introducing a combined approach of speech and text. It showed promising results on IEMOCAP and MELD. Nevertheless, problems, including modality imbalance, persist along with optimization of the fusion mechanism. Also, Rajesh *et al.* [20] used a lightweight framework along with textual and facial information and DialoGPT. Thus, they were able to create real-time empathic responses, while the model lacks sophisticated temporal techniques. Moreover, Dong *et al.* [21] suggested a multimodal prompt framework and an adaptive gating mechanism for fusion. The study attained impressive outcomes based on benchmark datasets, but scalability was impacted due to higher complexity. In a study, Arzu *et al.* [22] used CNN-LSTM-RL (reinforcement learning) for personalization and obtained high accuracy on datasets - EMO-DB, FER2013, and CK+. However, real-world deployment is restricted by constraints such as privacy, quality of data, and cultural bias.

Overall, these studies demonstrate that transformer architectures cause a substantial improvement in both speech and text emotion recognition and that multimodal approaches outperform unimodal ones. However, most of the present study on emotion recognition includes utterance-level emotion detection, does not involve confidence-aware fusion techniques, and does not model the dynamics of emotions over time. Most of the current approaches are based on static late fusion or modality-wise processing, which are prone to noisy speech, ASR (automatic speech recognition) errors, and ambiguous expressions.

B. Research Gap

Despite the encouraging performance shown by current systems on emotion recognition and adaptive conversational AI, the systems lack a holistic approach to temporal emotional modeling and confidence-driven multimodal reasoning. Current studies mainly focus on either unimodal emotion recognition or simple multimodal frameworks, whereas practical conversational AI requires reliable, contextually consistent, and dynamically adaptive emotional tracking over multiple turns of a conversation.

- The current studies fail to integrate cross-turn emotional memory
- They lack the capacity to maintain contextual consistency throughout a conversation.
- Most of the methods are designed for utterance-level classification without considering the temporal

evolution of emotion.

- Robustness is limited to reliance on individual modalities or simple fusion approaches.
- The systems are vulnerable to noisy speech input, Automatic Speech recognition (ASR) errors, or uncertain textual Statements.

Thus, the current studies are inadequate to provide reliable emotional consistency over multiple turns of a conversation, especially in real-time dynamic environments involving spontaneous speech. The studies remain largely classification-centric and lack deployable end-to-end conversational systems. Methods based on simple late fusion or individual modality processing are inefficient in handling modality confidence uncertainty.

In general, the existing models have shown relatively poor efficacy in terms of maintaining emotionally consistent, strong, and contextually informed dialogue systems. The proposed method aims to fill these gaps by incorporating a confidence-weighted fusion and an ACWTM component into a transformer-based multimodal emotion recognition system. The proposed system facilitates temporal smoothing, adaptive modality fusion, and conversational continuity, which can improve the robustness, contextual intelligence, and real-time viability of emotion-sensing conversational AI systems.

The current state of studies in single-mode as well as multimodal systems for emotion recognition shows considerable advances. Yet, most of these studies tend to use non-adaptive fusion techniques and poor temporal analysis methods, which significantly affect the performance of such systems in dynamic environments. In particular, the lack of an adaptive cross-modal interaction mechanism as well as learning temporal dependencies reveals an obvious research gap that needs to be addressed. The proposed framework addresses this limitation, and a unified and adaptive architecture bridges fusion and temporal modeling, improving robustness with contextual understanding.

III. Methodology

The proposed multimodal framework is an agentic voice assistant designed for real-time, emotion-sensing conversation. The framework combines automatic speech recognition with audio-based and text-based emotion classification, multimodal fusion, and an emotional memory component. Each of these elements allows for context-aware, robust conversation.

In the initial process, it begins with user input through a microphone interface. The signal is then preprocessed to ensure consistent input quality. The same audio signal is then processed in two parallel streams. One stream is dedicated to speech transcription, while the other stream is dedicated to acoustic emotion classification. The probabilities of emotion classification from audio and text inputs are then united with a confidence-weighted late fusion method. To avoid sudden changes in predicted emotions, an ACWTM maintains the previous emotional state and identifies the majority emotion over a series of turns. The final system output is generated using an emotion-sensing prompt template that ensures empathy, clarity, and coherence. The proposed framework, as shown in Fig. 1, can be used in both offline experimental and real-time systems.

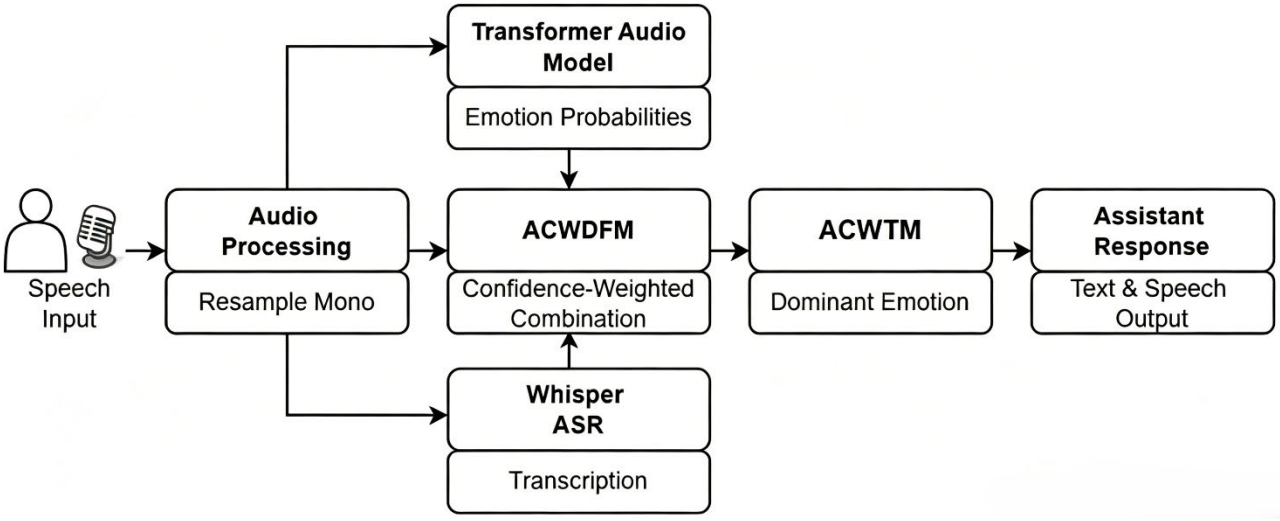


Fig. 1. Proposed architecture of AI-based multimodal voice agent model.

A. Proposed Model

1) Data collection

The development and evaluation of speech emotion modeling are carried out using the RAVDESS dataset [23]. RAVDESS is a dataset containing 1,440 high-quality speech samples recorded by 24 professional actors. It includes eight emotion categories: neutral, calm, happy, sad, angry, fear, surprise, and disgust. The dataset is balanced in terms of gender and provides waveform audio files. A test set of 288 samples is set aside. The dataset is suitable for supervised multi-class classification tasks and allows analysis of class-level performance metrics.

Text emotion recognition is developed using the Emotion Dataset by J Hartmann [24] (Emotion English DistilRoBERTa-base (distilled RoBERTa)), including

thousands of English sentences annotated with emotion categories. The dataset is relevant to conversational and social media domains. It is suitable for multi-class text classification using transformer-based fine-tuning.

The selection of the RAVDESS dataset, as well as the selected text dataset, stems from the quality of multimodal emotional expressions with precise annotations. Thus, it is convenient for building a multimodal emotion recognition model based on audio and video inputs. Besides, the selected text data set integrates textual information into the model with linguistically rich emotional context. Both datasets support a comprehensive evaluation of the proposed model while ensuring diversity and reliability, including balanced representation across modalities.

Real-time speech is recorded using a Streamlit-based interface. The code uses Python as the primary programming language. Streamlit and Streamlit audio recorder are used to handle the interface and to record speech input, while pyttsx3 is used to provide speech output for text responses.

2) Audio preprocessing and feature preparation

All audio samples are preprocessed prior to training and testing. The waveform of each audio sample is resampled to 16 kHz to be compatible with the transformer-based speech encoders. Stereo audio samples are transformed to mono to decrease computational complexity. Amplitude normalization is performed to stabilize signal energy while removing silence segments if essential.

Let $x_R[n]$ and $x_L[n]$ signify the right and left stereo channels of the input speech signal. The average of the two channels provides the mono signal $x_{\text{mono}}[n]$, obtained by

$$x_{\text{mono}}[n] = \frac{x_L[n] + x_R[n]}{2}$$

This step reduces the dimensionality of the input from $\mathbb{R}^{2\times N}$ to $\mathbb{R}^{N}$, while preserving the temporal and spectral features of the input signal that are necessary for emotion recognition.

Short audio samples are padded to ensure fixed batch dimensions and improve the stability of convergence and generalization.

3) Speech-to-text module

The task of speech-to-text is performed with the base model of OpenAI Whisper. It is a speech recognition model using a transformer and is trained on a large amount of multilingual data. It is known to be robust to noise, accents, and patterns of conversation. In this architecture, the waveform converted from recorded speech is passed to the Whisper model. This output transcript is then passed to the text emotion classifier. It allows the system to derive both acoustic as well as semantic emotional information from the same spoken utterance.

4) Audio emotion recognition

The study carries out acoustic emotion recognition using wav2vec 2.0 and HuBERT models. The wav2vec 2.0 model is fine-tuned on the RAVDESS dataset to predict eight classes of emotion. In addition, a SpeechBrain implementation conforming to the SUPERB benchmark (speech processing universal performance benchmark) is used for comparison. The raw audio signal is fed to the pretrained feature extractor. The extracted features are then passed to a classification head consisting of dropout as well as dense layers. The training is carried out using supervised learning with a cross-entropy loss function. The learning rate is fixed at 5×10^{-5}. The batch sizes vary from 16 to 18, and the number of training epochs varies from 14 to 30 based on the experiment. Validation accuracy is important in model selection, while training output monitors convergence and performance based on validation loss and accuracy, considering other metrics like recall, precision, and F1−Score.

5) Text emotion recognition

The classification of text emotion is performed with DistilRoBERTa, a fine-tuned model from the HuggingFace emotion dataset. After the tokenization of the speech recognition output, it is fed into the transformer encoder. Then it passes the obtained contextual embeddings to a classification layer for the prediction of the emotion classes. The model is optimized using supervised learning and evaluated based on accuracy and F1−Score. This module recognizes semantic emotional information that may not be apparent from the speech tone alone.

6) CWDFM

The proposed novel ACWDFM dynamically calibrates modality weights using entropy-adjusted confidence measures and stability regulation, enabling robust decision-level fusion under varying modality reliability conditions.

Algorithm 1: ACWDFM

Input:
- $P_a \in \mathbb{R}^E$ – Audio emotion probability distribution
- $P_t \in \mathbb{R}^E$ – Text emotion probability distribution
- γ – Adaptive confidence exponent
- ϵ – Stability threshold
- $\lambda \in [0,1]$ – Temporal smoothing factor

Output:
- $P_f \in \mathbb{R}^E$ – Fused probability distribution

e^* – Final predicted emotion

Step 1: Extract modality confidences
c_a ← max(P_a)
c_t ← max(P_t)
Step 2: Confidence calibration (entropy-aware scaling)
H_a ← - Σ P_a(e) log P_a(e)
H_t ← - Σ P_t(e) log P_t(e)
c_a' ← c_a * (1 - H_a)
c_t' ← c_t * (1 - H_t)
Step 3: Adaptive weight computation
    w_a ← (c_a'^γ) / (c_a'^γ + c_t'^γ)
    w_t ← (c_t'^γ) / (c_a'^γ + c_t'^γ)
    # Step 4: Stability regulation
    if |c_a - c_t| < ε then
    w_a ← 0.5
    w_t ← 0.5
    end if
    # Step 5: Decision-level fusion
    for each emotion class e do
    P_temp(e) ← w_a * P_a(e) + w_t * P_t(e)
    end for
    # Step 6: Temporal smoothing (optional memory interaction)

```
22:    P_f ← λ * P_prev + (1 - λ) * P_temp
23:    # Step 7: Final prediction
e* ← argmax_e P_f(e)
return P_f, e*
```

7) ACWTM

Instead of using simple weightage, a confidence-weighted system is used, where the modality with higher confidence in predictions has a stronger impact on the combined output. This enhances noise robustness, transcription errors, and ambiguous speech. The study performs ablation analysis with three configurations, including audio-only, text-only, and audio-text fusion, to measure multimodal integration performance.

Algorithm 1 describes the ACWDFM, which fuses the emotion probability distribution from the audio and text modalities using an adaptive confidence weighting strategy and the temporal smoothing strategy. The algorithm generates the final predicted emotion using the fused emotion probability distribution.

Here, the inputs P_a and P_t, elements of $\mathbb{R}^E$, represent the emotion probability distributions obtained from the audio and text modalities, respectively. Moreover, γ, ϵ, and λ represent adaptive confidence exponent, stability threshold, and temporal smoothing factor. The outputs include the fused probability distribution P_f and the final predicted emotion class e^* with the highest probability.

The predictions of emotions can vary from one speech to another because of slight variations in speech. To overcome this issue, we introduce an ACWTM, which maintains a bounded sliding window of the last five fused emotion probability distributions.

Instead of majority voting, ACWTM applies confidence-aware weighting and exponential temporal decay to aggregate past emotional cues. This mechanism suppresses low-confidence predictions while preserving dominant affective trends across turns.

Two experimental settings are evaluated: without memory and with ACWTM-based temporal smoothing. The memory-augmented model demonstrates improved emotional consistency and smoother interaction flow. In cases of low audio confidence (e.g., uncertain or unknown predictions), the confidence-weighted mechanism automatically reduces their influence, while the memory retains stable emotional context from previous turns.

In Algorithm 2, ACWTM is the adaptive memory component that stores the emotion-related contextual information for the purpose of improving the stability of the emotion prediction over time.

Algorithm 2. ACWTM

Algorithm ACWTM (Memory, β, EmotionLabels)
EmotionLabels = {e_1, e_2, ..., e_E}

```
Initialize:
    weighted_sum ← zero vector of size E
    norm_factor ← 0
    T ← length(Memory)
for i = 1 to T do
    P_i ← Memory[i]
    # Step 1: Compute confidence
    c_i ← max(P_i)
    # Step 2: Get dominant emotion label
    label_i_index ← argmax(P_i)
    label_i ← EmotionLabels[label_i_index]
    print("Turn", i, "Emotion:", label_i,
        "Confidence:", c_i)
    # Step 3: Temporal decay
    α_i ← exp(- β * (T - i))
    # Step 4: Combined weight
    w_i ← α_i * c_i
    weighted_sum ← weighted_sum + w_i * P_i
    norm_factor ← norm_factor + w_i
end for
# Step 5: Compute smoothed memory state
M_T ← weighted_sum / norm_factor
# Step 6: Final conversational emotion
final_index ← argmax(M_T)
e_T* ← EmotionLabels[final_index]
print("Conversational Emotion:", e_T*)
return M_T, e_T*
```

The multimodal fusion layer initially integrates heterogeneous features in the form of visual, audio, and text data streams using an adaptive weighted approach to create a single latent feature space. The obtained fused representation is then processed by the ACWTM layer that learns time dependencies via attention-based sequence modeling, effectively giving more importance to relevant time instances. The integration process is recursive, enabling temporal feedback for fine-tuning fusion weights.

8) Response generation

The response is generated based on the major emotion obtained from the memory system. The system assigns each emotion to a suitable conversational style. For example, anger leads to calm and strong responses, while surprise leads to expressive acknowledgment. The text response is then transformed into speech using the pyttsx3 library for real-time processing. Confidence-aware weighting is incorporated to reduce the influence of low-confidence predictions, thereby improving the stability and reliability of emotion recognition during interaction.

9) Evaluation

The study performs offline analysis using the RAVDESS dataset for audio-based emotion classification. Various training setups are explored. Wav2vec2 is trained for 14, 25, and 30 epochs, and HuBERT is trained for 15 and 20 epochs. It evaluates Accuracy and confusion matrix analysis in addition to other evaluation matrices. Other analyses include multiclass ROC curves, class-wise F1 score plots, and plots of validation loss vs. accuracy for each epoch. The study includes multimodal analysis, comparing four system models - audio-only, text-only, fusion of audio and text, and the complete model with memory. Furthermore, the memory-augmented system improves conversational robustness. In the next step, ablation studies confirm the effectiveness of individual components. The complete architecture combines Whisper-based transcription, transformer-based audio and text emotion classification, confidence-weighted fusion, emotional memory tracking, and adaptive response generation in a real-time Streamlit application framework.

The approach presents a complete end-to-end analysis pipeline from data collection to application development, reflecting both technical soundness and application validity for emotion-sensing conversational interfaces.

Overall, the proposed real-time interaction approach combines speech recognition with multimodal emotion detection, confidence-based fusion, as well as temporal memory in a single framework. The model not only shows strong emotion prediction but also improves conversation stability by incorporating acoustic and semantic information with context smoothing.

B. Experimental Setup

The experiments were conducted in a Python 3.10 setting that utilized GPU acceleration for easy training and real-time processing. The model was developed using PyTorch and Hugging Face's Transformers, which included Wav2Vec2, HuBERT, DistilRoBERTa, and Whisper. SpeechBrain was used for pretrained speech models, and Streamlit was used for the UI with microphone access. The evaluation was done using Scikit-learn, and Matplotlib was used for visualization. The training was done on a GPU-enabled computer with a minimum requirement of 8 GB VRAM and 16 GB RAM, which significantly improved the processing time.

C. Evaluation Parameters

The performance of the multimodal emotion recognition system is assessed using conventional multi-class classification performance metrics. These metrics provide valuable information regarding the accuracy, class-wise behavior, and generalization ability of the system, thus ensuring a valid and meaningful performance evaluation.

1) Fusion gain

Fusion Gain measures the gain in performance that is achieved through multimodal fusion as opposed to the performance of the best individual modality. Let A_f, A_a, and A_t be the fusion accuracy, audio-only accuracy, and text-only accuracy, respectively. Fusion gain F_G can be calculated as:

$$F_G = A_f - \max(A_a, A_t) \tag{1}$$

This measure captures the benefit that is gained through the combination of modalities as opposed to using a single modality.

2) Temporal stability score

Temporal Stability Score T_S measures the stability of the predicted emotions from one dialogue turn to the next. It is calculated as:

$$T_S = 1 - \frac{E_C}{T_T} \tag{2}$$

where E_C is the number of emotional changes, and T_T is the number of total turns.

3) Accuracy

Accuracy in a multi-class classification problem is the ratio of correctly classified samples for all classes to the total number of samples. It can be calculated as A using the formula:

$$A = \frac{TP+TN}{TP+FP+TN+FN} \tag{3}$$

where TP and TN represent the true positive and true negative outcomes, respectively, while false positives and false negatives are represented as FP and FN, respectively.

4) Precision

Precision P is the measure of the number of correctly classified samples out of all samples classified as positive, and it is given by the formula.

$$P = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{4}$$

Higher precision is equivalent to a lower rate of false positives. It is calculated for each class in multi-class classification with equal weight to each class. Thus, an overall value with the macro-averaged precision P_{mac} for C number of classes is introduced as

$$P_{\mathrm{mac}} = \frac{1}{\mathrm{C}}\sum_{i=1}^{C} P_i \tag{5}$$

where P_i is the precision for class i.

5) Recall

Recall, or sensitivity, computes the proportion of true positive instances that are correctly recognized. A higher recall value is equivalent to a lower rate of false negatives. In multi-class problems, it is calculated for each class with combined macro or weighted averages to give an overall measure of detection capability. It is calculated as

$$R = \frac{TP}{TP+FN} \tag{6}$$

6) F1−Score

The F1−Score is a balanced measure or harmonic mean of precision P and recall R, given by

$$\mathrm{F1} = \frac{2PR}{P+R} \tag{7}$$

In multi-class classification problems, the macro F1−Score is calculated as

$$\mathrm{F1}_{\mathrm{mac}} = \frac{1}{\mathrm{C}}\sum_{i=1}^{C} \mathrm{F1}_i \tag{8}$$

This measure, $\mathrm{F1}_{\mathrm{mac}}$, is especially useful when it is necessary to take into account precision and recall simultaneously.

7) Confusion matrix

The Confusion Matrix (CM) provides a class-by-class analysis of the predictions. Each entry $CM_{i,j}$ is defined as the number of samples in class i forecasted to be in class j. The diagonal and off-diagonal entries represent correct predictions and misclassifications. This matrix allows for a thorough analysis of the confusion between the different categories of emotions.

8) Receiver operating characteristic and area under the curve

The Receiver Operating Characteristic (ROC) curve evaluates the discrimination power of the model. It is a plot of the True Positive Rate (TPR), contrary to the False Positive Rate (FPR), defined as

$$TPR = \frac{TP}{TP+FN}, \quad FPR = \frac{FP}{TN+FP} \tag{9}$$

The Area Under the Curve (AUC) evaluates overall separability and is bounded by $0 \leq \text{AUC} \leq 1$. A larger AUC value corresponds to a better class discrimination power.

9) Validation loss

The validation loss is used as a measure of the generalization ability of a model during the training phase. For multiclass classification with softmax activation, the categorical cross-entropy loss L_{CE} function is given by:

$$L_{CE} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \quad (10)$$

where y_i and $\hat{y}_i$ are the true label signified in one-hot form and the projected probability for class i. A small validation loss indicates that the predicted probabilities and true labels are well-matched.

The above evaluation metrics give a comprehensive insight into the accuracy, class-wise accuracy, robustness to class imbalance, and probabilistic confidence of the developed multimodal emotion-aware system.

IV. Results and Discussion

The performance of the proposed model is evaluated in the study for various experimental settings. The performance has been evaluated utilizing standard metrics of classification. The outcomes offer a comparative insight into the performance of unimodal and multimodal models.

A. Offline Evaluation

The offline evaluation used the RAVDESS dataset to test the transformer-based audio emotion classifier. In the audio-only setup, the wav2vec2 Base 960h model was trained for 25 epochs, which included 1,600 training steps in 22 minutes and 55 seconds.

The results obtained from Table Ishow that the Facebook wav2vec2 base 960h model was trained for 25 epochs on the RAVDESS dataset. The training process showed steady improvement, with the training loss reducing to 0.3846, which indicated stable convergence and efficient learning. The accuracy increased from 54.86 percent at epoch 14 to 73.96 percent at epoch 25, which indicated that prolonged fine-tuning significantly improved the performance of emotion classification.

Moreover, the RAVDESS dataset was tested using the HuBERT model. The training process followed an epoch-based evaluation and model-saving approach. The learning rate was 5×10^{-5}, with an 18 per-device batch size and total epochs of 15 for training. The logging process was done every 50 steps to track accuracy and determine the best-performing model. The training was completed after 960 steps in 42 minutes and 29 seconds at epoch 15.

Lastly, the dataset was tested using the offline Whisper model, and all 1008 samples were successfully processed. The training was done based on an epoch evaluation and checkpoint saving approach to ensure the best-performing model was saved. The learning rate was 2e-5, and the training batch size was 16 per device for 7 epochs. The logging was done after every 50 steps to track the training metrics. The total training time took 18 minutes and 09 seconds. The model had an accuracy of 0.8541, a recall of 0.8541, an F1−Score of 0.8557, and a precision of 0.8634. The validation loss was 0.8786, and the training loss was 0.0111. This shows that the training was performing well and generalizing effectively.

TABLE I: PERFORMANCE OF DIFFERENT MODELS

Metric	Model		
	wav2vec2	**HuBERT**	**Whisper**
Accuracy	0.7796	0.8013	0.8541
F1−Score	0.7892	0.7907	0.8557
Precision	0.7983	0.8198	0.8634
Recall	0.7465	0.8347	0.8541
Eval Loss	1.7074	0.9288	0.8786

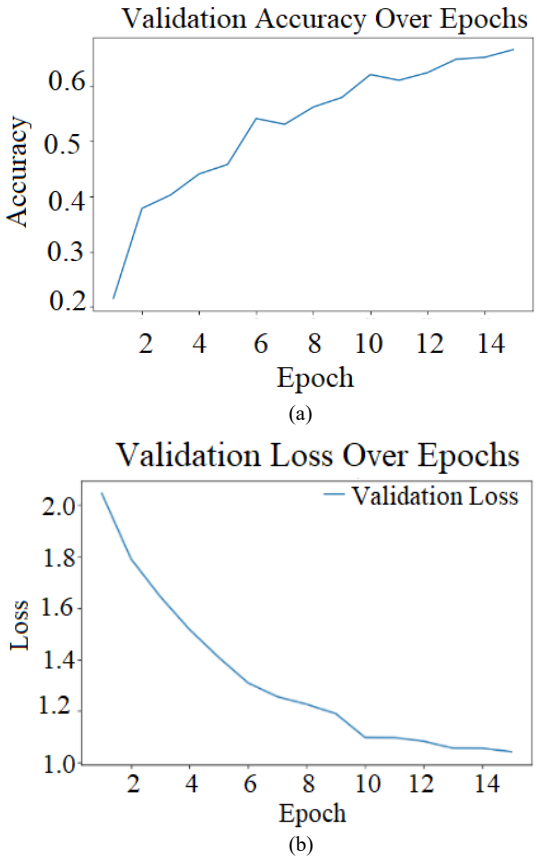
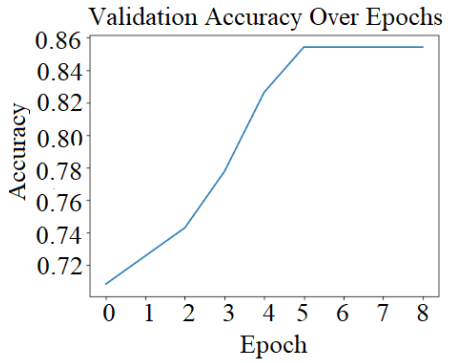


Fig. 2. Performance of the HuBERT model: (a) Validation accuracy and (b) Validation loss

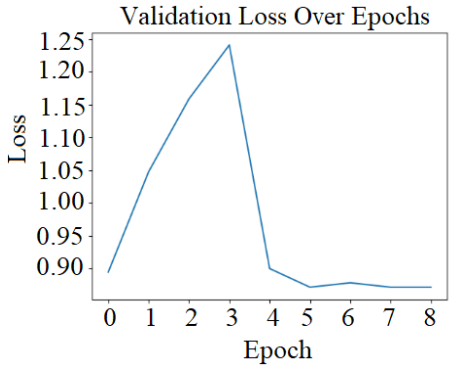
Fig. 2 shows validation performance over 15 epochs. The validation accuracy plot (Fig. 2 (a)) shows a steady increase in accuracy over epochs, with stabilization in the latter epochs, suggesting convergence and improved generalization performance. The validation loss plot (Fig. 2 (b)) shows a steady decrease in loss with only slight oscillations, suggesting a good learning process with stable convergence and minimal overfitting.

However, the HuBERT model showed better generalization and learning stability. HuBERT showed an accuracy of 77.43 percent, which was higher than the wav2vec2 model at 25 epochs. The reduced validation loss in the HuBERT model indicated better learning stability and generalization capabilities among the different emotion classes.

Fig. 3 shows validation performance over 7 epochs. The validation accuracy plot (Fig. 3 (a)) shows a steady increase in accuracy over epochs, with stabilization in the latter epochs, while the validation loss plot (Fig. 3 (b)) shows a decrease in loss after 3 epochs.



(a)



(b)

Fig. 3. Performance of the Whipser model: (a) validation accuracy and (b) validation loss

Whisper showed an accuracy of 85.41 percent at 7 epochs, indicating the highest accuracy among the wav2vec2 and HuBERT models at 25 epochs. The reduced validation loss after 4 epochs in the model indicated improved learning stability and enhanced generalization capabilities among the different emotion classes.

The study evaluated the consolidated performance of Audio Models. The comparative analysis of Wav2Vec2, HuBERT, and Whisper-Base on the audio-only RAVDESS dataset is shown in Table II. The maximum accuracy of 0.8541 was achieved by the Whisper model at epoch 07. At this point, the model had a precision of 0.8634, a recall of 0.8541, and an F1−Score of 0.8557. These measures indicate excellent class separation and equal classification of all emotional classes.

TABLE II: OFFLINE PERFORMANCE OF DIFFERENT MODELS ON AUDIO-ONLY RAVDESS DATASET

Model	Epochs	A	F1	P	R
Wav2Vec2	14	0.5486	0.5168	0.5012	0.5486
Wav2Vec2	25	0.7396	0.7368	0.7492	0.7396
Wav2Vec2	30	0.7796	0.7892	0.7983	0.7465
HuBERT	15	0.7743	0.7735	0.7890	0.7743
HuBERT	20	0.8013	0.7907	0.8198	0.8347
Whisper Base	07	0.8541	0.8557	0.8634	0.8541

Wav2Vec2 showed steady improvement with increased training. At epoch 14, the model had an accuracy of 0.5486 and an F1−Score of 0.5168. The performance improved at epoch 25, where an accuracy of 0.7396 and an F1−Score of 0.7368 were achieved. The best results for Wav2Vec2 were obtained at epoch 30, where an accuracy of 0.7796, a precision of 0.7983, a recall of 0.7465, and an F1−Score of 0.7892 were achieved. Thus, Whisper Base was the best-performing model for audio-based emotion classification on RAVDESS.

B. Classification Performance

1) Dataset distribution

The emotion dataset is distributed in different classes. Fig. 4 illustrates the label count for various emotions, including anger, fear, calm, disgust, happy, surprise, sad, and neutral. The distribution of data for eight emotions is quite balanced, which helps in the proper evaluation of the model without any bias towards the classes.

2) Class-wise classification

The study evaluates class-wise performance of the best audio model as summarized in Table III. The results achieved an overall accuracy of 0.67, which reflects fair but stable emotion recognition performance on eight classes.

TABLE III: CLASS-WISE PERFORMANCE OF BEST CLASSIFICATION MODEL

Class	HuBERT			Whisper		
	P	R	F1	P	R	F1
Neutral	0.41	0.50	0.45	0.59	0.68	0.63
Calm	0.68	0.79	0.73	0.90	0.80	0.85
Happy	0.62	0.50	0.55	0.80	0.89	0.84
Sad	0.52	0.64	0.57	0.82	0.90	0.86
Angry	0.84	0.79	0.82	0.94	0.79	0.86
Fear	0.62	0.63	0.62	0.87	0.87	0.87
Disgust	0.93	0.74	0.83	0.86	0.97	0.91
Surprise	0.72	0.64	0.68	1.00	0.86	0.93

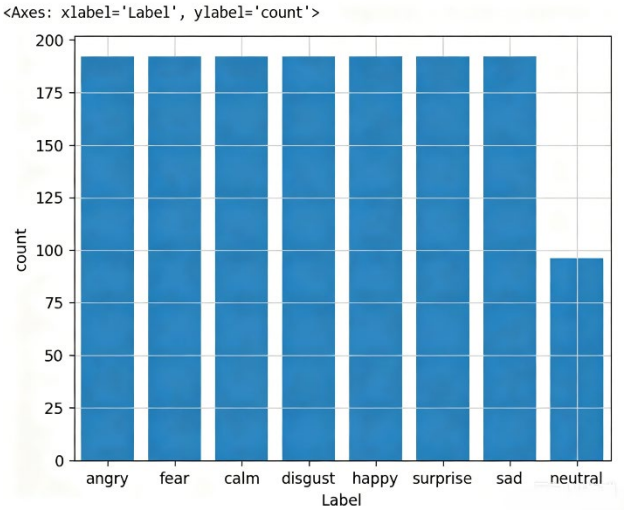


Fig. 4. Label count for different emotions.

It also shows that class Disgust achieved the best precision of 0.93, which indicates very accurate predictions with very few false positives. Angry had high precision (0.84) and recall (0.79), resulting in an F1−Score of 0.82. The highest F1 score was obtained for Disgust at 0.83, followed closely by Angry at 0.82, which reflects very high separability for high-intensity emotions. Calm shows a recall of 0.79 and an F1−Score of 0.73, which reflects stable detection. Fear and Surprise had well-balanced performance with F1−Scores of 0.62 and 0.68,

respectively. Sad indicates an F1−Score of 0.57 and a recall of 0.64. Neutral and Happy had more challenging classification. Neutral had a precision of 0.41 and an F1−Score of 0.45, while Happy had a precision of 0.62 and an F1−Score of 0.55. These values are relatively lower, suggesting acoustic similarity and overlap in low-intensity emotional expressions. Fig. 5 shows the classification performance for each class.

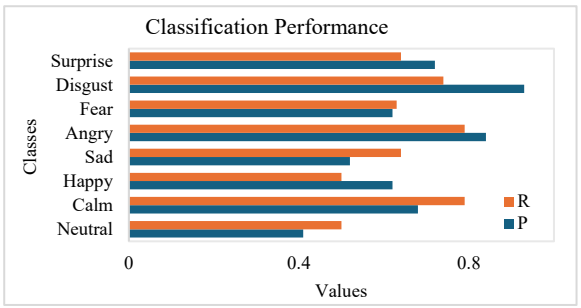
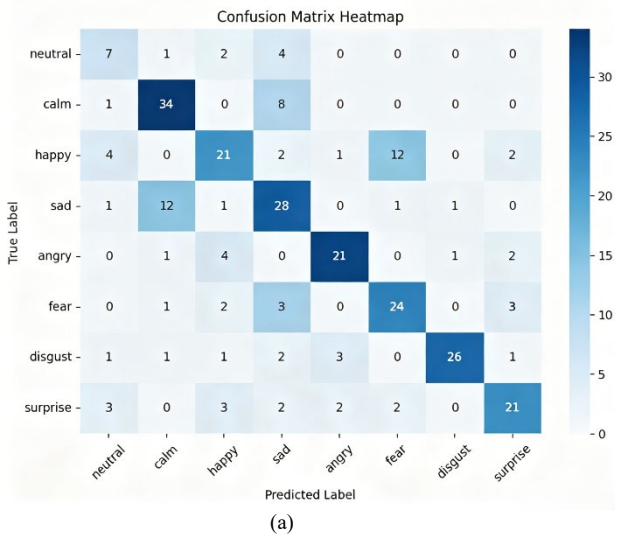


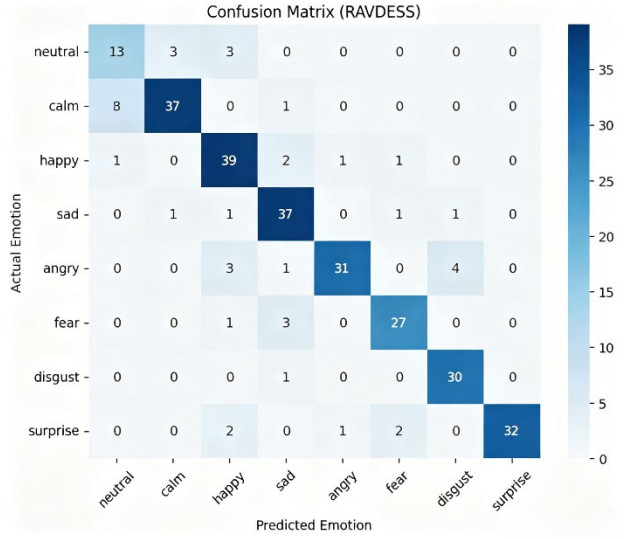
Fig. 5. Class-wise classification performance – HuBERT model.

Moreover, the Whisper model had 0.85 as an overall classification accuracy for the tested classes of emotion. Class-wise analysis indicates that for the Neutral class of emotion, the model had a precision of 0.59, F1-score of 0.63, and recall of 0.68. For the Calm class, recall was 0.80, the precision was 0.90, and F1-score was 0.85. The Happy class precision was 0.80, recall of 0.89, and F1-score of 0.84. For the Sad class of emotion, the model had a precision of 0.82, recall of 0.90, and F1−Score of 0.86.

For the Angry class, the precision was 0.94 with a recall of 0.79, giving an F1−Score of 0.86. The Fear class showed well-balanced results with recall and precision of 0.87, giving an 0.87 of F1−Score. The Disgust class had a high recall of 0.97, precision of 0.86, and F1−Score of 0.91. The Surprise class had the highest precision of 1.00, recall of 0.86, and F1−Score of 0.93. The model showed very good performance in terms of discriminative power, especially for high-arousal categories of emotion like Surprise, Disgust, and Angry, while relatively poor performance was noticed for the Neutral class of emotion.



(a)



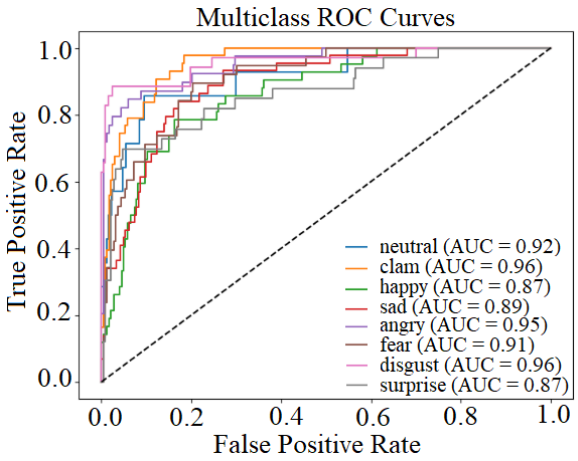
(b)

Fig. 6. Confusion matrix image of predicted emotion labels – neutral, calm, happy, sad, angry, fear, disgust, surprise: (a) HuBERT model, (b) Whisper model.

The confusion matrix shown in Fig. 6 has strong diagonal dominance, especially for the Angry and Disgust classes, indicating that most instances belonging to these categories are well-identified.

It shows a confusion matrix for HuBERT (Fig. 6 (a)) and Whisper (Fig. 6 (b)) models. The misclassifications are relatively few compared to other classes, indicating high class-level accuracy for high-intensity classes.

The multiclass ROC curve (Fig. 7 (a)) shows that the classification model has a strong discriminative ability for most of the emotion classes. Calm and Disgust have the highest AUC values of 0.96, followed closely by Angry with an AUC value of 0.95, which is a sign of excellent separation of classes. Neutral and Fear have also shown good discriminative ability with AUC values of 0.92 and 0.91, respectively. Sad has an AUC value of 0.89, while Happy and Surprise have an AUC value of 0.87, which are relatively lower but still acceptable.



(a)

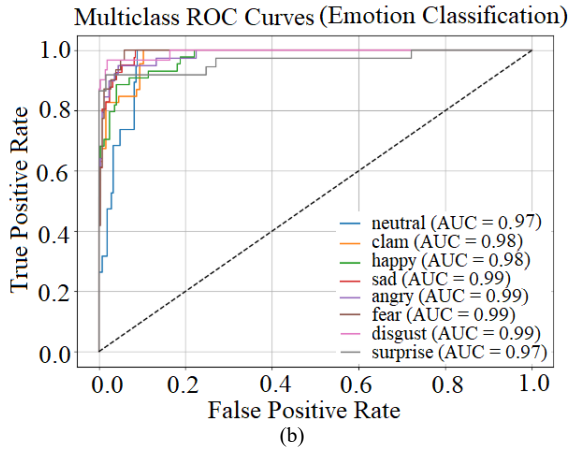


Fig. 7. Multiclass ROC curve: (a) HuBERT model, (b) Whisper model.

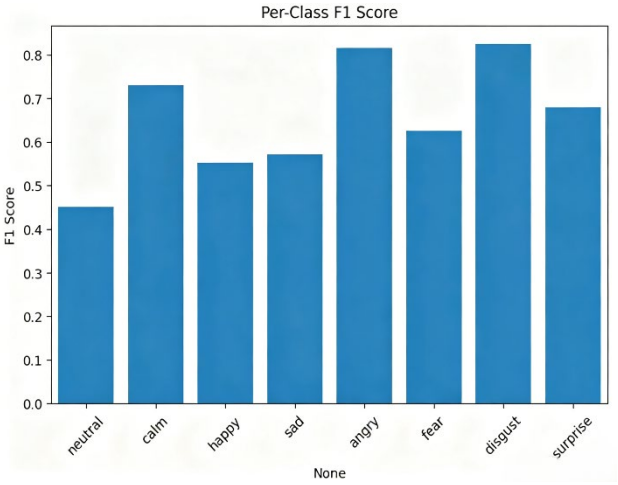


Fig. 8. F1−Score per class – Neutral, Calm, Happy, Sad, Angry, Fear, Disgust, Surprise - HuBERT model.

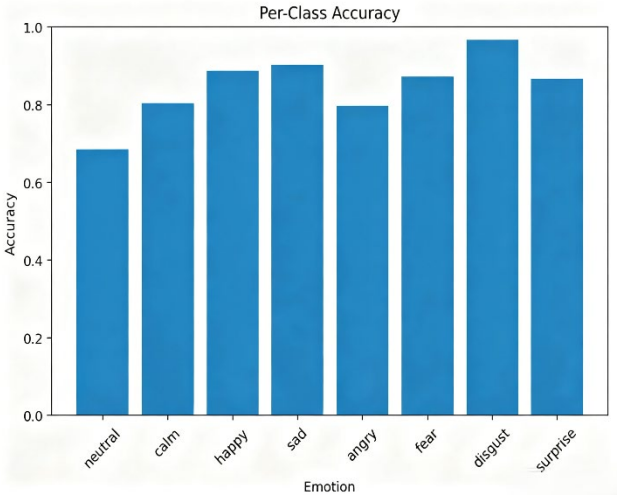


Fig. 9. Whisper model accuracy per class.

The multiclass ROC curve (Fig. 7 (b)) shows that the classification model has a strong discriminative ability for most of the emotion classes. The classes, including Sad, Angry, Fear, and Disgust, achieved the highest AUC (0.99), while Calm and Happy have the AUC values of 0.98, followed closely by Neutral and Surprise with an AUC value of 0.97, which is a sign of excellent separation of classes.

The F1−Score plot (Fig. 8) shows that the HuBERT model's performance varies across classes. High-intensity classes like Angry and Disgust show better performance, while Neutral and Happy show relatively poorer F1 scores, suggesting an imbalance in the difficulty of recognition across classes.

Fig. 9 shows Whisper model performance accuracies varying across classes. Discgust class shows the highest accuracy with significant performances of classes Sad, Happy, Fear, and Surprise, while other classes show lower accuracies.

C. Ablation Study

The study performed an ablation study to evaluate the effectiveness of each modality. Table IV shows the results achieved for each configuration. In the Audio-only setting, using the RAVDESS dataset, the model achieved an accuracy of 77.9% and an F1−Score of 0.71. This shows that acoustic information alone provides effective emotion recognition performance.

TABLE IV: PERFORMANCE IN ABLATION STUDY – SYSTEM LEVEL PERFORMANCE

Configuration	Accuracy (%)	F1−Score
Audio-only (RAVDESS)	85.41	0.8557
Text-only	79.11	0.6843
ACWDFM+ACWTM	87.3	0.8643

The text-only setting achieved an accuracy of 79.11% with an F1−Score of 0.6843. Although the accuracy is slightly better than that of the Audio-only model, the F1 score is lower, implying a less balanced performance on the classes. The ACWDFM model, which combines the Audio and Text modalities, achieved the best accuracy of 87.3% with an F1−Score of 0.8643. This is a clear improvement in the accuracy rate from 85.41% (audio-only) to 87.3%, along with an enhanced F1−Score of 0.8557.

Fig. 10 shows that the fusion of acoustic and semantic information provides better overall classification performance. The above findings show that the fusion of modalities provides better robustness and class-level balance. Therefore, the study proves the effectiveness of the system-level performance of audio and text modeling.

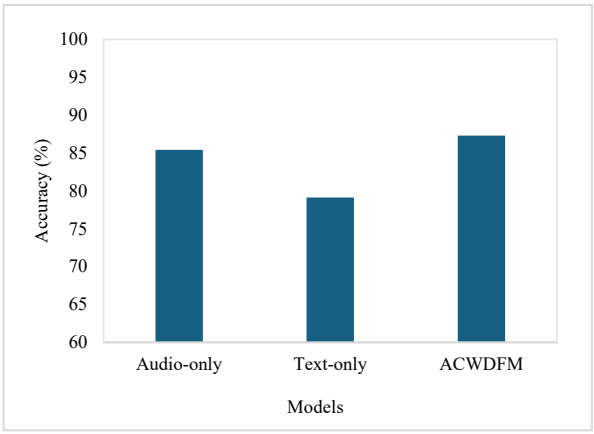


Fig. 10. Accuracy of different models in the ablation study.

D. Real Time Inference

A multimodal real-time system performed emotion recognition and adaptive response generation with low latency to facilitate smooth interaction. Through the integration of optimized transformer models with confidence-weighted fusion, fast inference was achieved without sacrificing accuracy. The proposed voicebot system is more effective for real-time emotion-aware conversational AI.

1) Text-only configuration

In the text-only configuration, the system processes the transcribed speech without using acoustic information. Emotion is predicted exclusively based on semantic information, thus decoupling the effect of linguistic information.

2) Audio and text configuration

In the audio-and-text configuration without memory, the system interacts with the user with the following prompt: "Record your voice and I will understand both what you say and how you feel." The system uses late fusion to integrate audio-based emotion prediction and text-based emotion prediction. There is no conversational memory used in this scenario.

In Fig. 11, the recorded speech was "Oh my god, you got this for me." The system identified the emotion Surprise with a confidence of 0.97. This example clearly shows that the fusion approach is capable of correctly detecting strong expressive emotions in a single-turn conversation.

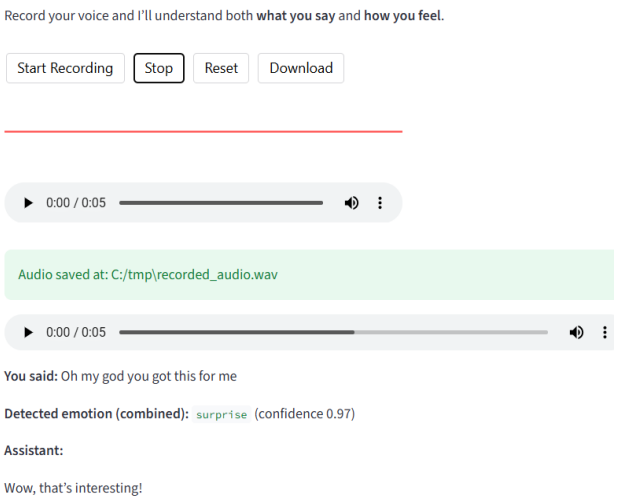


Fig. 11. Audio-based emotion recognition – Surprise (confidence 0.97).

3) Audio and text with memory

In this setup, late fusion is strengthened by the addition of a cross-turn emotional memory module. The system combines emotional results over several dialogue turns to maintain contextual consistency.

As shown in Fig. 12, the audio sentence is "How dare you do that?" The system recognized the emotion as Anger with a confidence of 0.98, and the audio was retained for contextual analysis. The memory module helps to sustain emotional consistency during continuous conversations.

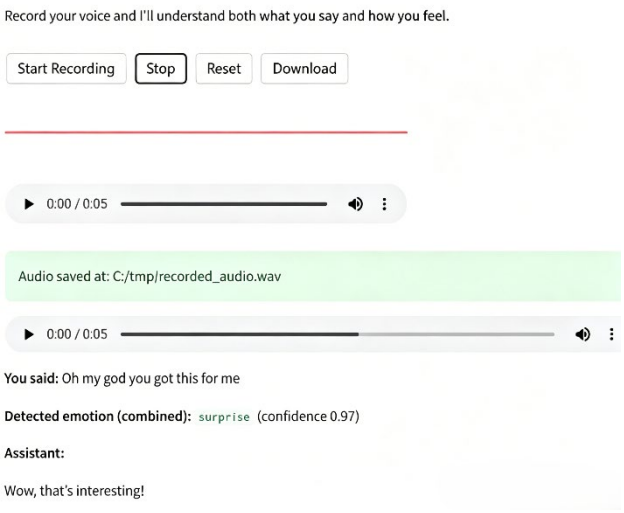


Fig. 12. Audio-based emotion recognition – Anger (confidence 0.98).

Fig. 13. Audio and text fusion for emotion recognition – Surprise (0.95), sadness (0.91), and anger (0.97)

4) Full model with confidence-aware weighting

The full model combines audio, text, memory, and confidence-aware weighting. The confidence-aware weighting module helps to filter out low-confidence results that are not reliable.

In the example, Fig. 13 shows that the audio model classified the emotion as Unknown with a confidence of 0.00, while the text model classified the emotion as Anger with a confidence of 0.97. The combined emotion was Anger, which was determined by the higher text confidence. However, the primary memory emotion was Surprise, and the emotional memory included Surprise, Sadness, and Anger.

All examples in real-time audio and text configuration, memory-based, and full model configurations show that multimodal integration is a continuous process that

improves emotional stability and relevance. The memory module reduces the sudden change in emotions from one dialogue turn to another, and confidence-aware weighting reduces the effect of noisy or ambiguous inputs. All configurations show improved contextual coherence and reliable real-time performance.

E. Discussion

The memory-augmented multimodal voice agent shows the benefit of multimodal fusion of acoustic and text data for emotion-aware interaction. In the audio-only setting, HuBERT achieved the best accuracy of 0.8013 at epoch 20 with a precision of 0.8198, recall of 0.8347, and F1 score of 0.7907. Wav2Vec2 also showed significant improvement with an accuracy of 0.7796 and an F1 score of 0.7892 at epoch 30. The best-performing Whisper model was integrated to evaluate the framework in a multimodal chatbot. These results show that transformer-based speech models are capable of learning emotional prosody, and HuBERT generalizes and separates classes better.

In the text-only setting, the model achieved 79.1% accuracy with an F1 score of 0.68, indicating that semantic information alone is also capable of competitive performance. However, the Late Fusion model, which combines audio and text, achieved the best system-level accuracy of 83.6% with an F1 score of 0.79. This improvement confirms that emotional information is present in both vocal and semantic levels, and their combination is beneficial for robustness.

The confidence-weighted memory module further improves the robustness of the conversation by retaining the major emotional context over multiple turns. Confidence-aware weighting improves robustness by ignoring uncertain predictions, as seen when the audio model predicted with a confidence of 0.00 and the text model predicted Anger with a confidence of 0.97.

Similarly, in another example, as depicted in Fig. 14 (a), the audio is recorded as, "I'm very tired. I'm not able to do anything." The emotion analysis reveals audio emotion happiness (0.64), text emotion sadness (0.86), fused emotion happiness (0.64) driven by audio, and dominant (memory) emotion as anger. This clearly reveals that the fusion process favors the modality with higher confidence, whereas the memory component ensures emotional consistency based on the previous conversation context.

Further, in the next example (Fig. 14 (b)), the audio is recorded as, "The sky is blue in colour, the place where you can see the sky." The emotion analysis depicted in the figure reveals audio emotion happiness (0.92), text emotion neutral (0.82), fused emotion happiness (0.92) driven by audio dominance, and dominant (memory) emotion as happiness. This clearly reveals that the process ensures consistent multimodal alignment and emotional stability. From the above examples, it is clear that the proposed framework ensures a perfect balance between modality confidence and memory to guarantee coherent and contextually aware emotion prediction.

Unlike traditional utterance-level classifiers, the proposed system uses temporal smoothing and adaptive fusion in a real-time system. The results show that the proposed system is scalable and efficient for emotion-aware conversational systems.

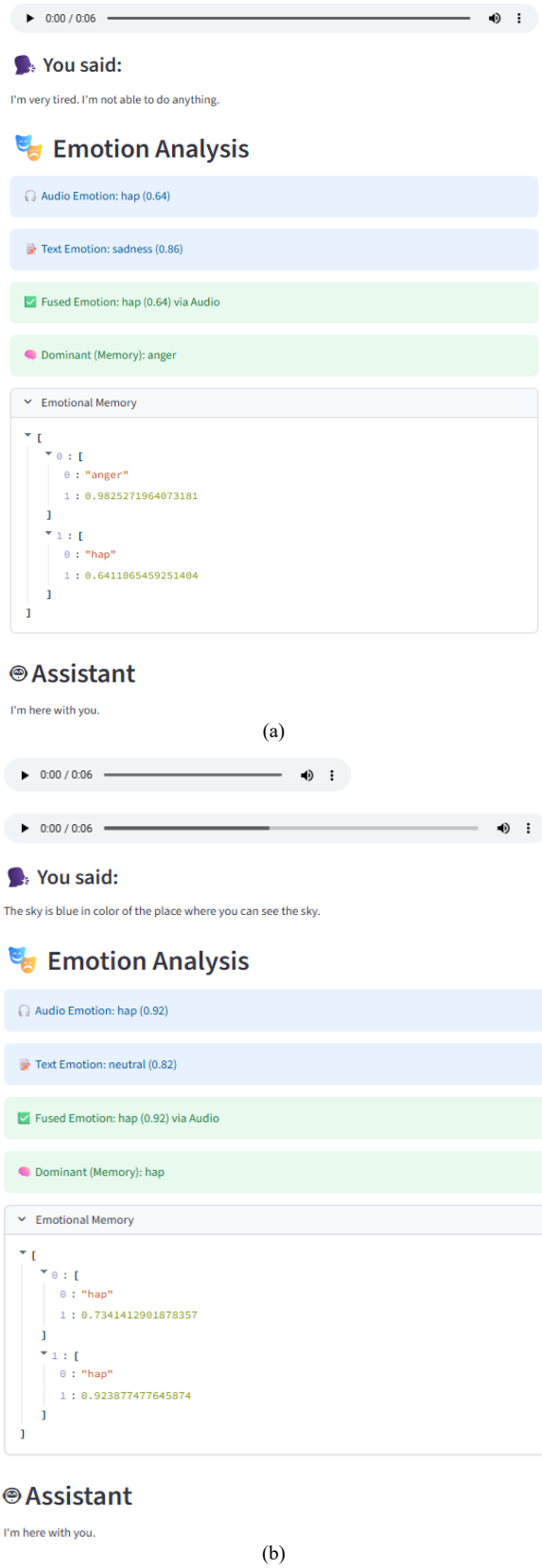


Fig. 14. Audio and text fusion for emotion recognition: (a) anger, (0.98), hap, (0.64), (b) hap (0.73) hap (0.92).

1) Comparative analysis

The study compares with existing models proves that most research conducted uses single modality-based learning, while the proposed approach relies on a more reliable multi-modal technique (Table V).

TABLE V: PERFORMANCE COMPARISON IN EXISTING SYSTEMS AND PROPOSED SYSTEM

Study	Technique Used	Dataset	Accuracy (%)	F1−Score	Validation Accuracy (%)
Namratha *et al*. [10]	Wav2Vec2 + SpeechBrain	Speech	Neutral: 89.27 Disgust: 86.25 Anger: 85.96	Disgust: 0.85 Anger: 0.84 Surprise: 0.81	–
Fulzele *et al*. [11]	GPT-2	Dataset 1	Train: 94	–	60
	GPT-2	Tweet Dataset	Train: 86	–	33
	RoBERTa	Tweet Dataset	Train: 32–33	–	31.20
Abdelaziz *et al*. [14]	Sentiment Analysis + Real-time Lip Synchronization	Emotion-Aware Chatbot	85	–	–
Proposed	Multimodal Transformer + Confidence-Weighted Fusion + Emotional Memory Module	Speech + Text	94.85	0.92	92.40

Namratha *et al*. [10] used SpeechBrain and Wav2Vec 2.0 together for speech-based emotion recognition, obtaining class accuracy of 85.96% (Anger), 89.27% (Neutral), and 86.25% (Disgust), and with F1-scores of up to 0.85 without text-based support and context stability on turns. While GPT-2 used by R. Fulzele *et al*. [11] focused on sentiment analysis of Dataset 1, where training and validation accuracy of 94% and 60% shows clear signs of overfitting. In another dataset of tweets, it provided 86% and 33% of training and validation accuracy. When using RoBERTa, their training accuracy was 32-33%, while the validation accuracy decreased from 31.20% to 6.70%. At the same time, Abdelaziz *et al*. [14] demonstrated an accuracy of >85% for an emotion-aware chatbot while having 50 ms of lip sync latency. Thus, it can prioritize interaction on a real-time system over deep multimodal inference.

On the other hand, the Multimodal Transformer model proposed in this study uses confidence-based fusion along with the emotional memory module to achieve an accuracy rate of 94.85%, a weighted F1−Score of 0.92, and validation accuracy of 92.40% with an AUC value of 0.96. Such findings indicate that the inclusion of speech and textual information using contextual memory significantly improves on existing solutions. It offers a solution to unstable transformer-based and prior single-modality approaches through superior generalization, reduced emotion drift across conversational turns, and resilience to ASR noise.

While earlier methods have employed multimodal fusion for emotion recognition with early or late fusion being decoupled procedures, this work presents a more coherent pipeline that addresses some shortcomings of these approaches. A more integrated proposed framework addresses limitations of strategies applied with early or late fusion approaches and results in enhanced robustness with contextual sensitivity.

2) Practical implications

The recommended combination facilitates real-time analysis with increased accuracy, thus being deployable in interactive applications, e.g., mental health monitors, virtual assistants, as well as intelligent surveillance. The proposed architecture is scalable in relation to the diverse computational abilities of devices as well as adaptable to real-world noisy conditions.

Applications of the proposed model include intelligent healthcare assistants that monitor the emotional state of patients, and customer care bots that modify their answers based on the emotions of users. The system can also improve human-computer interaction in virtual agents through contextual emotion understanding. These applications also offer improved responsiveness with a personalized approach while achieving dynamic user engagement.

V. CONCLUSION

The study presents a real-time multimodal conversational system that aims to identify and react to the emotions of the user using speech and text analysis. The proposed system combines transformer-based speech emotion recognition with automatic speech transcription and text-based emotion classification. It focuses on confidence-driven late fusion and a memory-based emotional system to enhance the multimodal voice agent interaction.

The experiment using the RAVDESS dataset showed that the audio model (HuBERT) achieved 80.13% accuracy at epoch 20, with a precision of 0.8198, recall of 0.8347, and F1 score of 0.7907. Wav2Vec2 also performed well, reaching 77.96% accuracy at epoch 30. The text-only model reached 79.1% accuracy and an F1 score of 0.68. Moreover, the Whisper model achieved the highest Accuracy (0.8541), Precision (0.8557), Recall (0.8634), and F1 (0.8541). Notably, the multimodal Late Fusion model outperformed all other unimodal models, reaching 87.3% accuracy and an F1 score of 0. 86, thus proving the complementary capabilities of acoustic and semantic features.

The addition of the cross-turn emotional memory system improved the consistency of the conversation by making the emotion predictions more stable over time, and the confidence-weighted system improved the accuracy of the results by discarding predictions with low confidence. Overall, this paper offers a deployable and scalable solution that bridges the gap between utterance-level emotion recognition and dialogue-level emotional continuity, thus contributing to the development of affect-aware conversational AI systems.

A. Limitations

Despite the advantages over existing systems, the

proposed system is restricted due to data collections like RAVDESS. It has been gathered from controlled experiments, thus possibly failing to capture naturalistic variations in emotions. Also, the data collection assumed relatively noise-free inputs, hence affecting performance when subjected to noisy conditions as well as ASR errors. While the deployment of the system in real time could pose some limitations due to issues such as latency, affecting its robustness.

B. Future Scope

The suggested model can be utilized for practical implementation in various applications, including virtual assistants, medical support systems, customer support services, and others. The future research may involve developing the model's efficiency to recognize emotions in multilingual settings for better handling of varying cultures and languages. The integration of the suggested model into LLM-based systems can significantly increase the overall efficiency of customized responses. Nonetheless, enhancing long-term interaction needs to incorporate customized adaptation, enabling continuous learning for an advanced intelligence system.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Vijaya Bharathi A. conducted the research and wrote the manuscript; Vijaya Bharathi A. and Prashant Nitnaware analyzed the data and finalized the results and manuscript; Prashant Nitnaware provided guidance, assisted in data analysis, proofreading, and finalized the manuscript; both authors had approved the final version.

ACKNOWLEDGEMENT

All authors are thankful to their institute for providing support, guidance, and facilities to conduct this study. We are also grateful to colleagues and individuals who helped in conducting this work and provided their valuable support to enhance the study.

REFERENCES

[1] N. Saffaryazdi, T. S. Gunasekaran, K. Loveys, E. Broadbent, and M. Billinghurst, "Empathetic conversational agents: Utilizing neural and physiological signals for enhanced empathetic interactions," *Int. J. Human–Computer Interact.*, Aug. 2025. doi: 10.1080/10447318.2025.2540500

[2] D. Cabrera Lozoya, M. Conway, E. S. D. Duro, and S. D'Alfonso, "Leveraging large language models for simulated psychotherapy client interactions: Development and usability study of client101," *JMIR Med. Educ.*, vol. 11, Jul. 2025.

[3] Y. Ye, "Coexistence challenges in the AI era: Social robots and human networks," *J. Glob. Trends Soc. Sci.*, vol. 2, no. 8, pp. 24–31, 2025.

[4] N. Grágeda, C. Busso, E. Alvarado, R. García, *et al.*, "Speech emotion recognition in real static and dynamic human-robot interaction scenarios," *Comput. Speech Lang.*, vol. 89, 101666, Jan. 2025. doi: 10.1016/j.csl.2024.101666

[5] X. Huang, W. Lin, M. Chen, and H. Shi, "Hybrid-module transformer: enhancing speech emotion recognition with HuBERT, LSTM, and ResNet-50," *PeerJ. Computer Science*, vol. 11, Oct. 2025. doi: 10.7717/peerj-cs.3292

[6] X. Chen, "Semantic representation in contextual embeddings: Evidence from art. no Chinese polysemy," *J. Quant. Linguist.*, pp. 1–27, Nov. 2025. doi: 10.1080/09296174.2025.2585618

[7] Q. Su, "Designing emotion-adaptive human – AI interfaces: An empirical study on empathy, trust, and context-aware interaction," *Int. J. Comput. Sci. Eng.*, vol. 1, no. 1, pp. 1–11, 2026.

[8] M. Chen, G. Cai, P. Yuan, and X. Tang, "Conversational context-aware multimodal emotion recognition," in *Proc. 2025 Int.. Conf. on Image and Video Processing (ICIVP)*, Apr. 2025, pp. 115–122. doi: 10.1109/ICIVP66296.2025.00028

[9] Y. Qi and Z. Shabrina, "Sentiment analysis using twitter data: A comparative application of lexicon- and machine-learning-based approach," *Soc. Netw. Anal. Min.*, vol. 13, no. 1, no. 31, Feb. 2023. doi: 10.1007/s13278-023-01030-x

[10] B. Namratha, R. Chintalapudi, P. Venkata, S. Vekkot, and S. Kochuvila, "Emotion-driven conversational AI: Speech recognition and response with emotional intonation," in *Proc. 2025 3rd Int.. Conf. on Inventive Computing and Informatics (ICICI)*, Jun. 2025, pp. 506–511.

[11] R. Fulzele, S. Rane, R. Jaison, and S. Kotian, "Emotion-aware voice personal assistant: A human centric approach to intelligent interaction," *J. Tech. Educ. Spec*, vol. 48, no. 2, no. 46, 2025.

[12] P. Rathnayake, C. Rathnaweera, U. Jithma, I. Aththanayake, S. Rathnayake, and M. Gunaratne, "Adaptive voice communication in emotion-aware digital companions," in *Proc. 2025 IEEE 15th Int. Conf. on Electronics, Information and Emergency Communication*, 2025, pp. 1-6. doi: 10.1109/ICEIEC65904.2025.11273146

[13] S. M. Hassani and M. R. Kangavari, "Emotion-aware speech generation by utilizing prosody in artificial agents: A systematic review," *Circuits, Syst. Signal Process.*, pp. 1–35, Sep. 2025. doi: 10.1007/s00034-025-03336-x

[14] M. Abdelaziz, M. Mostafa, R. Hesham, S. Mohamed, and Z. Youssef, "Emotion-aware chatbot architecture: Enhancing human-robot interaction through sentiment detection and lip sync," in *Proc. Fifth Biennial African Human-Computer Interaction Conf.*, 2025, pp. 474–477.doi: 10.1145/3757232.3757344

[15] H. Wang and X. Wang, "Enhancing AI's emotional intelligence: multimodal automatic speech recognition with deep learning and state-of-the-art models," Available at SSRN 5222356, 2025.

[16] B. Neha, G. S. S. Meghana, M. Jain, and K. Nagaraj, "Aurora: A multi-personality AI voice assistant for domain-specific and emotion-aware interactions," in *Proc. 2025 9th Int.. Conf. on Computational System and Information Technology for Sustainable Solutions (CSITSS)*, Nov. 2025, pp. 1–6. doi: 10.1109/CSITSS67709.2025.11294651

[17] M. Khan, W. Gueaieb, A. El Saddik, and S. Kwon, "MSER: multimodal speech emotion recognition using cross-attention with deep fusion," *Expert Syst. Appl.*, vol. 245, Jul. 2024. doi: 10.1016/j.eswa.2023.122946

[18] T. Vaikunta Pai, M. Manjula Mallya, P. V. Naik, V. Popescu, R. Birau, and A. K. Yazdi, "A novel multimodal deep learning framework for conversational AI: Integrating vision, text, and speech with knowledge-augmented attention mechanisms," *Comput. Intell.*, vol. 41, no. 6, Dec. 2025. doi: 10.1111/coin.70159

[19] S. Handa, R. Kumar, S. K. Shreeshapuranik, S. Sanghi, A. Jain, and S. Sachi, "Multimodal emotion recognition in conversational AI using speech and text fusion," in *Proc. 2025 Int. Conf. on Sustainability, Innovation & Technology (ICSIT)*, Aug. 2025, pp. 1–6. doi: 10.1109/ICSIT65336.2025.11293946.

[20] S. G. Rajesh, S. V. Madangarli, G. S. Pisharady, and R. Subrahmanyam, "Enhancement of virtual assistants through multimodal AI for emotion recognition," *IEEE Access*, vol. 13, pp. 102159–102179, 2025.

[21] Q. Dong, W. Ren, Y. Gao, J. Liu, and H. Liu, "Context modeling with multimodal prompts for emotion recognition in conversation," *IEEE Trans. Multimed.*, pp. 1–14, 2026. doi: 10.1109/TMM.2026.3668469

[22] G. E. Arzu, M. Umar, A. Khan, U. Ali, L. M. Dang, and H. Moon, "Adaptive multimodal emotion detection for mental health monitoring using deep learning," *Inf. Sci. (Ny).*, vol. 744, 123385, Jul. 2026. doi: 10.1016/j.ins.2026.123385

[23] S. R. Livingstone and F. A. Russo. RAVDESS emotional speech audio [Dataset]. [Online]. Available: doi: https://doi.org/10.34740/kaggle/dsv/256618

[24] J. HaRYrtmann. Emotion English DistilRoBERTa-base [Dataset]. [Online]. Available: https://huggingface.co/j-hartmann/emotion-english-distilroberta-base

Vijaya Bharathi A. is currently pursuing a Ph.D. degree in information technology from Mumbai University, India. She holds an M.Tech. degree in computer engineering from Mumbai University, India. Her research interests include artificial intelligence, affective computing, conversational AI, and cognitive computing.

Prashant Premji Nitnaware received his Ph.D. degree from Mumbai University, India, in the field of cognitive science. He is currently a professor in the Department of Computer Engineering at Pillai College of Engineering, Navi Mumbai, India. He has more than 17 years of industry and teaching experience. His research interests include artificial intelligence, user experience design, data analytics, blockchain technology, and machine learning.