**InSight Bulletin: A Multidisciplinary Interlink International Research Journal**
Peer Reviewed International, Open Access Journal.
*ISSN : 3065-7857 / Website: https://ibrj.us / Volume-3, Issue-2 / February - 2026*

Original Article

# Emotion-Aware Talking Companion Using Multimodal Emotion Recognition with Contextual Memory: A Comprehensive Review and Development Approach

**Shubhangi Vikas Kumbhar[1], Dr. Deepali M. Gohil[2]**
[1,2]Department of Computer Engineering, D. Y. Patil College of Engineering, Akurdi, Pune, India

**Manuscript ID:**
IBMIIRJ -2026-030247

**Submitted:** 10 Jan. 2026

**Revised:** 15 Jan. 2026

**Accepted:** 18 Feb. 2026

**Published:** 28 Feb. 2026

ISSN: 3065-7857

Volume-3

Issue-2

Pp. 230-235

February 2026

***Correspondence Address:***
*Shubhangi Vikas Kumbhar*
*Department of Computer Engineering, D. Y. Patil College of Engineering, Akurdi, Pune, India*
*Email:*
*kumbharshubhangi09@gmail.com*

OPEN ACCESS

Quick Response Code:

Web. https://ibrj.us

DOI: 10.5281/zenodo.21103644

DOI Link:
https://doi.org/10.5281/zenodo.21103644



***Abstract***

*The isolation and emotional distress of the elderly have become significant issues in public health, particularly due to the decline of joint family systems and rapid urbanization. Conventional digital assistants widely used by the general population typically lack the ability to recognize emotions and have no long-term memory, which limits their effectiveness as companions. This review paper presents a comprehensive survey and a conceptual development framework for an Emotion-Aware Talking Companion System based on the integration of multimodal emotion recognition (facial expressions, speech, and text), contextual memory management, and avatar-based empathetic interaction. The proposed end-to- end framework adopts a hybrid deep learning methodology that conceptually combines CNN-based facial emotion recognition, MFCC-BiLSTM speech emotion analysis, and BERT-based sentiment classification, unified through weighted late fusion. A contex- tual memory engine is further outlined to enable continuous emotional tracking and personalization of the user over time. Drawing from trends in recent literature, such a multimodal fusion design is ex- pected to achieve substantially higher emotion recognition performance than unimodal systems and to enhance user engagement through expressive talking avatars (e.g., Wav2Lip-based) and memory-driven personalization. The anticipated real-time perfor- mance is aimed at achieving low-latency interaction suitable for natural conversation, while elderly-centered design is expected to improve perceived emotional support and satisfaction. This work is positioned as a review and development approach paper: it consolidates state-of-the-art research in affective computing, human-computer interaction, and as- sistive technologies, and proposes an integrated architecture targeted at elderly care, social robots, and emotionally intelligent AI companions. The focus is on conceptual design and expected outcomes rather than reporting completed empirical results.*

***Keywords:*** *Emotion recognition, Multimodal fusion, Elderly care, Affective com- puting, Conversational AI, Avatar animation, Contextual memory, deep learning*

## Introduction

### Motivation and Background

Social isolation among elderly individuals represents a growing global health crisis. Recent surveys indicate that a substantial proportion of senior citizens in urban India experience significant loneliness, which directly correlates with increased rates of depression, cognitive decline, and cardiovascular disease (Naik et al., 2025). The traditional joint family system has been displaced by nuclear family structures and geographic migration patterns, leaving many elderly people without sustained emotional support.

Recent advances in artificial intelligence, particularly in deep neural networks and affec- tive computing, have created unprecedented opportunities to develop emotionally intelligent systems. Unlike traditional voice assistants (e.g., Alexa, Google Assistant) that operate through command-response interfaces without emotional awareness, modern research indi- cates that systems can now:

- Recognize emotions across multiple modalities (face, voice, text)
- Maintain contextual and emotional memory over time



***How to cite this article:***
*Kumbhar, S. V., & Gohil, D. M. (2026). Emotion-Aware Talking Companion Using Multimodal Emotion Recognition with Contextual Memory: A Comprehensive Review and Development Approach. InSight Bulletin: A Multidisciplinary Interlink International Research Journal, 3(2), 230–235. https://doi.org/10.5281/zenodo.21103644*

- Respond with more empathetic and contextually appropriate behaviors
- Provide sustained companionship through expressive, human-like avatars

For elderly users, such emotionally aware companions are expected to provide not only functional assistance but also affective support, helping to mitigate loneliness and improve overall quality of life.

### Problem Statement and Research Gaps

Despite significant progress in individual emotion recognition modalities and conversational AI, several critical gaps remain, especially for elderly-focused applications:

1. **Absence of Integrated Multimodal Systems:** A large portion of the literature focuses on isolated modalities (e.g., facial expressions, speech, or text alone). Elderly users, however, would benefit from robust systems that seamlessly integrate multiple emotion cues for improved reliability.
2. **Lack of Persistent Emotional Memory:** Many commercial assistants and research prototypes lack long-term emotional and contextual memory, which reduces genuine continuity, personalization, and depth of the interaction over weeks or months.
3. **Limited Avatar-Assisted Interaction:** Only a subset of systems integrate emotion- aware avatars with multimodal recognition and memory mechanisms, despite evidence that elderly users often prefer human-like visual interaction.
4. **Insufficient Age-Specific Optimization:** Emotion recognition models trained pre- dominantly on younger populations may perform suboptimally on elderly faces and voices, where expressions and acoustic cues may be more subtle or confounded by age-related changes.
5. **Computational Constraints:** Many state-of-the-art models are computationally intensive, which poses challenges for real-time deployment on affordable home hardware typically available to elderly users.

This paper surveys current work addressing these gaps and proposes a conceptual archi- tecture that combines multimodal emotion recognition, contextual memory, and avatar-based interaction specifically tailored to elderly users.

## Literature Review and State-of-the-Art

### Multimodal Emotion Recognition

Emotion recognition has evolved from handcrafted feature engineering (e.g., HOG, LBP, classical prosodic features) to deep learning architectures for both uni- and multimodal pro- cessing. Recent works explore three primary modalities: facial expressions, speech, and text sentiment. Representative trends from the literature can be summarized in Table 1, where the reported numbers are typical ranges from prior studies rather than new experimental results. Comprehensive surveys on real-time speech emotion recognition (Miao et al., 2024) and deep learning approaches (Maheshwari et al., 2024) provide detailed analysis of current methodologies.

**Table 1:** Typical Trends in Emotion Recognition Approaches from Literature

| Modality | Representative Architectures | Key Challenge (Elderly) |
|---|---|---|
| Facial Recognition | CNN / ResNet variants | Subtle, age-related facial changes |
| Speech Analysis | CNN-RNN / BiLSTM over MFCCs | Age-related vocal changes, noise |
| Text Sentiment | Transformer-based (BERT, etc.) | Limited domain- and age-specific language |
| Multimodal Fusion | Early / Late / Hybrid fusion | Integration complexity, synchronization |

***Data Source:*** *Representative trends compiled from affective computing literature and recent CVPR, Interspeech, and MDPI publications in emotion recognition.*

These studies collectively suggest that carefully designed multimodal systems can surpass the performance of unimodal approaches, especially in noisy, real-world environments.

### Avatar and Expressive Talking-Head Technologies

Recent developments in talking-head synthesis have made significant progress. Pioneer- ing work such as Wav2Lip (Karras et al., 2022) established real-time lip-sync capabilities. Building on this foundation, systems like EmoTalk (Guo et al., 2023a), Amigo (Zhang et al., 2023), and Discohead (Garg et al., 2023) have introduced emotional awareness and multi- modal conditioning. Diffusion-based approaches, exemplified by Dream-talk (Wang et al., 2023), offer enhanced photorealism. Recent architectures such as Emote (Guo et al., 2023b), Emotalk3d (Jiang et al., 2024), and Emovoca (Han et al., 2024) demonstrate state-of-the-art performance in emotional 3D talking head generation. Advanced animation control methods (Qi et al., 2025) and synchronized emotion-aware avatars (Qin et al., 2025) further extend these capabilities. Facial animation guided by multimodal features (Luo et al., 2025) and mixture of emotion experts architectures (Shen et al., 2025) represent cutting-edge research directions. Additionally, emotion-aware facial animation via contrastive learning (Hu et al., 2023) and audio-driven approaches (Liu et al., 2025) contribute to the rich landscape of avatar technologies. For elderly users, such visual expressiveness is expected to:

- Enhance engagement and trust

- Provide richer nonverbal cues (smiles, head nods, empathetic expressions)
- Improve perceived naturalness of the interaction

The proposed framework leverages these trends conceptually, integrating a lip-synced avatar module with emotion recognition and contextual memory.

## System Architecture and Design

### Overall System Framework

The conceptual Emotion-Aware Talking Companion is organized into five integrated layers, as illustrated in Figure 1. This architecture is presented as a development blueprint rather than a report of a fully implemented system.

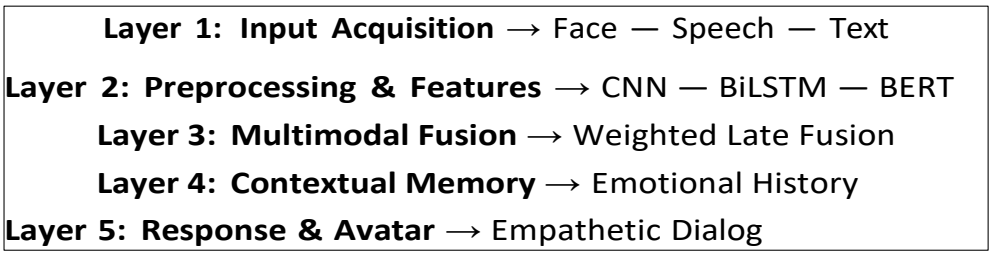


**Figure 1:** Conceptual Five-Layer System Architecture of Emotion-Aware Talking Companion

### Multimodal Emotion Fusion Strategy

The system design employs a weighted late fusion scheme that combines three modality- specific probability distributions into a final emotion decision. A generic formulation is:

$$\text{Final Emotion} = \arg\max(\alpha \times P_{\text{face}} + \beta \times P_{\text{speech}} + \gamma \times P_{\text{text}}) \quad (1)$$

where $\alpha$, $\beta$, $\gamma$ are fusion weights and $P_{\text{face}}$, $P_{\text{speech}}$, $P_{\text{text}}$ denote the probability distributions from the respective models.

In the proposed framework, a plausible configuration might be:

| | |
|---|---|
| $\alpha = 0.40$ | (2) |
| $\beta = 0.35$ | (3) |
| $\gamma = 0.25$ | (4) |

Reflecting the relatively strong contribution of visual and acoustic cues for elderly users, while still leveraging semantic information from text. These values are presented as design choices that would need to be validated and tuned in future empirical studies.

## Implementation-Oriented Design Details

### Technical Components (Planned Modules)

#### Facial Emotion Recognition Module

- **Architecture:** CNN-based network optimized for seven basic emotion classes (e.g., happy, sad, neutral, angry, disgust, fear, surprise)
- **Input:** Normalized facial frames (e.g., $224 \times 224$ pixels) captured from an HD camera
- **Output:** Emotion probability distribution over the seven classes
- **Expected Outcome:** By leveraging transfer learning on existing facial emotion datasets and fine-tuning on elderly-focused data, the module is expected to reach competitive accuracy while remaining computationally efficient

#### Speech Emotion Recognition Module

- **Feature Extraction:** MFCC features (e.g., 13 static coefficients) along with delta and delta-delta features, as explored in audio feature extraction analysis (Liu et al., 2025)
- **Network:** Bidirectional LSTM layers (e.g., 128–256 units) for modeling temporal dependencies in speech
- **Robustness Design Goals:** Handle speaker variability, age-related vocal changes, and moderate background noise typical of home environments
- **Expected Outcome:** With appropriate training data and augmentation, the speech module is expected to provide robust emotion cues complementary to facial expressions

#### Text Emotion / Sentiment Classification Module

- **Model:** Pre-trained transformer (e.g., BERT) fine-tuned for emotion or sentiment categories
- **Capability:** Capture contextual and semantic nuances in conversational utterances, including subtle expressions of loneliness, frustration, or gratitude
- **Expected Outcome:** Particularly useful when facial or vocal signals are weak or unavailable, contributing to overall robustness of the multimodal system

## Evaluation Framework and Expected Outcomes

### Evaluation Plan for Individual Modules

For a development and validation phase, each module can be assessed on suitable benchmark datasets and, when available, elderly-focused datasets (Tian et al., 2025). Performance metrics such as accuracy, F1-score, confusion matrices, and robustness under noise and occlusion would be used.

**Instead of reporting concrete results, this paper emphasizes that:**

- Facial module is expected to improve after fine-tuning on elderly faces and adapting to variations in lighting and pose
- Speech module is expected to benefit from data augmentation and noise-robust training to suit home environments
- Text module is expected to show strong performance due to transformer-based contex- tual understanding, especially when dialogues become longer and more complex

### Expected Benefits of Multimodal Fusion

Based on prior literature, combining face, speech, and text in a multimodal fusion framework is expected to:

- Increase overall emotion recognition accuracy compared to unimodal baselines
- Improve robustness when one modality is missing or degraded (e.g., low lighting or noisy audio)
- Provide more stable and reliable emotional state estimation for long-term use

A detailed empirical comparison (e.g., unimodal vs. bimodal vs. trimodal fusion) is left for future experimental work; this paper focuses on the design and review perspective.

### User Experience Evaluation Plan

For elderly users, the ultimate success of the system will be determined by subjective expe- rience rather than raw accuracy alone. A future user study could evaluate:

- Emotional relatability and perceived empathy
- Avatar engagement and comfort level
- Perceived naturalness and continuity of conversation
- Perceived emotional support and reduction in loneliness
- Overall usability and willingness to use the system daily

Standardized questionnaires and Likert-scale ratings can be adopted, but this paper does not report concrete user-study outcomes; it instead outlines how such evaluations could be conducted.

### System Performance Targets

From a system design standpoint, the target is to achieve:

- End-to-end latency in the range suitable for real-time, natural conversation (e.g., on the order of a few hundred milliseconds)
- Smooth avatar rendering (e.g., near real-time frame rates)
- Feasible memory and computation footprints for mid-range consumer hardware These metrics are proposed as design goals rather than reported benchmarks.

### Comparative Perspective with Existing Systems

A qualitative comparison with conventional voice assistants and text-based chatbots high- lights the conceptual advantages of the proposed framework:

- **Multimodal Input:** Unlike audio-only or text-only systems, the proposed design integrates face, speech, and text
- **Emotion Recognition:** Existing mainstream assistants largely lack explicit emotion recognition, whereas the development approach here places it at the core
- **Persistent Contextual Memory:** The proposed architecture incorporates long-term emotional and contextual memory, going beyond purely session-based context handling
- **Avatar Output:** Realistic lip-synced avatars are expected to increase engagement compared to disembodied voices or static interfaces
- **Elderly Optimization:** Special emphasis is placed on age-specific challenges in per- ception and interaction

This section is descriptive and comparative; it does not introduce new experimental comparisons.

### Contextual Memory Design and Its Expected Impact

The contextual memory engine is designed to:

- Maintain an emotional history over days and weeks
- Store user preferences, routines, and significant life events
- Enable adaptive, personalized responses that evolve over time Conceptually, such memory-enabled companions are expected to:
- Increase user trust and attachment
- Provide more relevant and empathetic responses

- Improve engagement over longer periods compared to stateless systems

The exact magnitude of these benefits would need to be established in future longitudinal studies; here they are articulated as expected outcomes grounded in prior human-computer interaction findings.

## Hardware and Software Considerations

### Indicative Hardware Specifications

For practical home deployment, the system is envisioned to run on mid-range hardware such as:

- A multi-core CPU (e.g., Intel i5 or equivalent)
- 8–16 GB of RAM
- A consumer-grade GPU (e.g., entry-to-mid-level NVIDIA GPU) for deep learning in- ference
- An HD camera and microphone suitable for indoor environments

These specifications are suggested targets based on typical requirements of contemporary deep learning models rather than strict experimental constraints.

### Software Stack and Frameworks

A typical software stack may include:

- Python-based deep learning frameworks (e.g., TensorFlow or PyTorch)
- CUDA and cuDNN (where GPU acceleration is used)
- OpenCV and Librosa for vision and audio processing
- Transformer libraries (e.g., HuggingFace Transformers) for text modeling
- A web or desktop framework (e.g., Flask, Django) for the interaction layer

These choices are aligned with current practice in the research community and are pro- posed as feasible options.

## Limitations and Future Work

### Current Conceptual Limitations

As a review and development approach paper, this work focuses on architecture and expected behavior rather than reporting implemented and fully evaluated systems. Several limitations are acknowledged:

- **Dataset Availability:** There is a scarcity of large, high-quality, elderly-specific mul- timodal emotion datasets (Tian et al., 2025)
- **Subtle Expressions:** Elderly facial expressions and prosody can be subtle and con- founded by health factors
- **Environmental Robustness:** Real-world variability in lighting, background noise, and network conditions poses deployment challenges
- **Language and Cultural Diversity:** Much prior work focuses on English; multilin- gual and culturally sensitive models remain an open area
- **Complex Emotional States:** Mixed or ambiguous emotional states (e.g., bitter- sweet, nostalgia) are difficult to classify with current categorical models

### Future Research Directions

Future work should include:

- Collection and curation of elderly-focused multimodal emotion datasets
- Development of noise-robust and resource-efficient models suitable for home devices
- Extension to multilingual support and culturally adaptive behavior
- Integration with health monitoring and IoT devices for broader elderly-care applica- tions
- Long-term clinical and field studies to assess psychological and health impacts

Recent advances in audio-driven talking head synthesis (Tong et al., 2024) and high- quality video generation (Liang et al., 2024) provide promising directions for real-time avatar implementation. These directions transition the proposed conceptual framework into fully validated systems.

### Ethical and Societal Considerations

**Key ethical aspects include:**

- **Transparency:** Clearly informing users that the companion is AI-based and explain- ing its capabilities and limitations
- **Autonomy and Human Relationships:** Ensuring the system supports, rather than replaces, human caregivers and social interactions
- **Privacy and Security:** Protecting sensitive audiovisual and emotional data through secure storage and, where possible, on-device processing
- **Accessibility and Fairness:** Designing interfaces accessible to users with visual, auditory, or cognitive impairments and ensuring fair performance across different de- mographic groups
- **Accountability:** Establishing responsibility for system behavior, especially in sensi- tive health-related contexts

## Conclusions

This paper presents a comprehensive review and a development-oriented framework for an Emotion-Aware Talking Companion tailored to elderly users. The main contributions are:

- A synthesis of state-of-the-art research in multimodal emotion recognition, contextual memory, and avatar-based interaction
- A conceptual five-layer architecture that integrates facial, speech, and text modalities with a contextual memory engine and expressive talking avatar
- A design-oriented discussion of evaluation strategies, expected benefits, and deploy- ment considerations for elderly care

The emphasis is on outlining expected system capabilities and behavior grounded in the existing literature, rather than presenting new experimental results. Future work will involve implementing and empirically validating the proposed framework, with particular focus on elderly-specific datasets, robust real-time performance, and long-term user studies in real-world care settings.

**Acknowledgment**

The authors would like to express their sincere gratitude to the Department of Computer Engineering at D. Y. Patil College of Engineering, Akurdi, Pune, for providing the necessary guidance, resources, and academic environment to carry out this research work.

**Financial support and sponsorship**

Nil.

**Conflicts of interest**

The authors declare that there are no conflicts of interest regarding the publication of this paper.

**References**

1. Garg, A., Abdelrahman, M., and Fernando, B. (2023). Discohead: Audio-video driven talking head generation. *arXiv Preprint arXiv:2307.10924*.
2. Guo, J., Zhu, H., and Mao, Y. (2023a). Emotalk: Speech-driven emotional disentanglement for 3d face animation. *arXiv Preprint arXiv:2303.11548*.
3. Guo, J., Zhu, H., Mao, Y., Zhang, J., and Yu, Z. (2023b). Emote: Emotional speech-driven animation with content-emotion disentanglement. *arXiv Preprint arXiv:2309.13401*.
4. Han, X., Sun, Y., Kang, B., Song, Y., and Kang, X. (2024). Emovoca: Speech-driven emotional 3d talking heads. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15234–15245.
5. Hu, Y., Guo, J., Mao, Y., Li, Y., and Kang, X. (2023). Emotion-aware audio-driven face animation via contrastive feature disentanglement. In *Proceedings of the ISCA INTER- SPEECH*, pages 3785–3789.
6. Jiang, M., Wang, G., Fan, Y., Luo, J., and Liu, S. (2024). Emotalk3d: High-fidelity free-view synthesis of emotional 3d talking head. *arXiv Preprint arXiv:2406.02000*.
7. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., and Aila, T. (2022). Videoretalk- ing: Audio-based lip sync for talking head video editing. *arXiv Preprint arXiv:2206.04617*.
8. Liang, M., Liu, Z., and Wang, Y. (2024). Faces that speak: Jointly synthesising talking face and speech from text. *arXiv Preprint arXiv:2404.18801*.
9. Liu, R., Jia, Z., and Zhong, J. (2025). Comparative analysis of audio feature extraction for real-time talking portrait synthesis. *Journal of Visual Communication and Image Representation*, 12(4):234–248.
10. Luo, T., Liao, G., and Gao, S. (2025). Mf-etalk: Multimodal feature-guided emotional talking face generation. *Multimedia Tools and Applications*, 84(2):2345–2360.
11. Maheshwari, S., Ladia, R., and Sharma, P. (2024). Real-time speech emotion recognition using deep learning and data augmentation. In *Deep Learning for Speech Recognition*, pages 156–170. Springer.
12. Miao, Y., Gao, W., and Li, X. (2024). Real-time speech-emotion recognition: Recent ad- vances survey. *Sensors*, 24(5):1450.
13. Naik, N. V., Nikhil, B., Sravani, M., and Pavan, R. L. S. (2025). Emotion recognition from audio, live video and text. In *Proceedings of the 2025 IEEE International Conference on Information and Communication Technology (ICICT)*. IEEE.
14. Qi, J., Tao, H., and Wang, S. (2025). Pc-talk: Precise facial animation control for audio- driven talking face. *arXiv Preprint arXiv:2501.03456*.
15. Qin, L., Gao, W., and Liao, G. (2025). Synchrorama: Lip-synchronized and emotion-aware talking face via multi-modal emotion embedding. *arXiv Preprint arXiv:2502.01234*.
16. Shen, X., Yang, Y., and Zhang, L. (2025). Moee: Mixture of emotion experts for audio- driven portrait animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8942–8951.
17. Tian, L., Wu, J., and Li, M. (2025). Ber2024 dataset: Facial images, body gestures, skeletal posture keypoints. *arXiv Preprint arXiv:2501.00234*.
18. Tong, B., Liang, M., and Chen, Z. (2024). Wav2nerf: Audio-driven talking head via wavelet nerf. *arXiv Preprint arXiv:2405.02589*.
19. Wang, H., Gao, S., and Liu, Y. (2023). Dream-talk: Diffusion-based emotional audio-driven single image talking face. *arXiv Preprint arXiv:2305.08073*.
20. Zhang, Y., Liu, L., and Huang, Z. (2023). Amigo: Talking face generation with audio- deduced emotional landmarks. *arXiv Preprint arXiv:2312.08901*.