Asian Journal of
Science, Technology, Engineering, and Art

e-ISSN: 3025-4507
p-ISSN: 3025-5287
Index: Harvard, Boston, Sydney University, Dimensions, Lens, Scilit, Semantic, Google, etc

https://doi.org/10.58578/AJSTEA.v4i2.8617

Multimodal Conversational AI: A Review of Integration Techniques, Applications, and Future Directions

Herbert Wanga
University of Iringa, Tanzania
wangahp@gmail.com

Article Info:

Submitted:	Revised:	Accepted:	Published:
Jan 1, 2026	Feb 25, 2026	Mar 13, 2026	Mar 18, 2026

Abstract

The integration of multimodal inputs, including text, voice, and visual data, into conversational artificial intelligence (AI) systems marks a significant shift toward more natural and effective human–computer interaction. This narrative synthesis review examines recent research on the technological foundations, applications, challenges, and future directions of multimodal conversational AI. Drawing on prior studies, the review analyzes key frameworks and models, including Situated Interactive MultiModal Conversations (SIMMC) and DialogueTRM, which employ multimodal fusion to support emotion recognition and context-aware interaction. The synthesis indicates that combining multiple modalities enhances system accuracy, strengthens user engagement, and enables richer contextual understanding in conversational settings. At the same time, the review identifies major challenges related to data synchronization, privacy protection, computational complexity, and bias mitigation. Based on these findings, the study highlights the need for future research on adaptive fusion techniques, cross-cultural usability, ethical AI development, and the incorporation of emerging modalities such as haptic and physiological data. This review contributes to the growing scholarship on conversational AI by providing an integrated understanding of the opportunities

and limitations of multimodal systems and by outlining directions for the development of more responsive, inclusive, and ethically grounded AI interactions.

Keywords: Conversational Artificial Intelligence; Human–Computer Interaction; Multimodal Fusion; Multimodal Interaction; Narrative Synthesis Review

Introduction

Conversational AI has evolved from text-based chatbots to systems capable of processing multimodal inputs, including voice, facial expressions, gestures, and, increasingly, haptic and physiological signals (McTear, 2017; Cassell, 2001). This shift aims to emulate human-like communication, where meaning is constructed from multiple sensory and contextual channels. Virtual assistants like Siri, Alexa, and Google Assistant now incorporate visual and auditory cues to disambiguate user intent (Feldman et al., 2017). Despite rapid progress, research remains fragmented across modalities, and a cohesive understanding of effective integration strategies, ethical implications, and real-world scalability is needed. This review aims to consolidate current knowledge, identify gaps, and suggest pathways for future development in multimodal conversational AI.

This paper is structured around four key themes:

1. Technological Foundations: Advances in automatic speech recognition (ASR), computer vision, natural language processing (NLP), and multimodal fusion architectures.

2. Applications: Case studies in healthcare, education, embodied conversational agents (ECAs), and customer service.

3. Challenges: Data fusion, synchronization, privacy, computational demands, bias, and real-world deployment barriers.

4. Future Directions: Adaptive systems, cross-cultural design, ethical frameworks, and the inclusion of emerging input types.

Multimodal Inputs in Conversational AI

Text and Voice Integration

Text and voice are foundational modalities for conversational AI. Modern systems use ASR and NLP to interpret user intent, but voice-only interactions often lack nuanced contextual awareness (McTear, 2017). Combining text and voice inputs enables redundancy and error correction. For example, Feldman et al. (2017) demonstrated that parallel processing of speech and transcribed text improves accuracy by cross-referencing modalities, especially in noisy environments or for users with atypical speech patterns.

Visual and Gestural Inputs

Visual data, such as facial expressions, eye gaze, and body language, provides affective and contextual cues that text or voice alone cannot capture. Systems like the Rea embodied agent (Cassell, 2001) and CareAdvisor (Feldman et al., 2017) use cameras and sensors to track user emotions, enabling more empathetic and contextually appropriate responses. For instance, detecting a user's confused expression or affirmative nod can dynamically adjust the AI's dialogue strategy or pacing.

Emerging Modalities: Haptic and Physiological Data

Beyond the traditional triad, research is exploring the integration of haptic feedback (e.g., touch, force) and physiological signals (e.g., heart rate, galvanic skin response) to create even richer interactive experiences. These modalities can convey emotional arousal, stress levels, or user engagement in real-time, offering potential applications in telemedicine, virtual reality, and advanced driver-assistance systems (Khan Mohd et al., 2022).

Multimodal Fusion Techniques

Effective integration of multimodal data requires robust fusion strategies, each with distinct trade-offs between complexity and interpretability.

Table 1. Comparison of Multimodal Fusion Techniques

Technique	Description	Advantages	Disadvantages
Early Fusion	Combines raw or low-level features (e.g., audio spectrograms + image pixels) for joint processing (Potamianos et al., 2003).	Can model complex inter-modal relationships early; potentially higher performance.	Sensitive to noise and misalignment; requires synchronized data; less interpretable.
Late Fusion	Merges decisions or predictions from unimodal models (e.g., averaging intent probabilities from separate text and voice classifiers; Korbar et al., 2018).	Robust to missing modalities; uses state-of-the-art unimodal models; easier to debug.	Cannot capture low-level cross-modal interactions; may miss synergistic cues.
Hybrid Fusion	Balances early and late interactions, often using attention mechanisms or intermediate representations (e.g., Factor Graph Attention; Schwartz et al., 2020).	Flexible; can capture interactions at multiple levels.	Architecturally complex; requires careful design and more data.

Case Studies

Situated Interactive MultiModal Conversations (SIMMC)

Moon et al. (2020) introduced SIMMC, a framework for task-oriented dialogues grounded in shared visual environments (e.g., virtual reality shopping). Key innovations include:

- Dynamic Context Updating: The assistant updates visual scenes in real-time based on user inputs (e.g., "rotate that chair").

- Fine-Grained Visual Grounding: Dialog acts and coreferences are explicitly linked to visual objects, effectively resolving ambiguities like "the brown one on the left."

DialogueTRM for Emotion Recognition in Conversation (ERC)

Mao et al. (2020) proposed DialogueTRM, a transformer-based model that captures intra- and inter-modal emotional cues. Its architecture features:

- Hierarchical Transformer (HT): Separately models context-dependent (textual) and context-free (visual/audio) inputs.

- Multi-Grained Interactive Fusion (MGIF): Aligns modalities at both neuron and vector levels, reported to improve ERC accuracy by 10.4% on the MELD benchmark.

Comparative Insight:

While SIMMC excels in visually-grounded task completion, DialogueTRM focuses on affective understanding in social dialogue. This highlights the modality specialization required for different application domains.

Challenges and Future Directions

Challenges

i. Technical Hurdles: Precise data synchronization remains difficult, especially for real-time streams like speech and gesture (Cassell, 2001). Computational complexity for real-time multimodal processing is a barrier to deployment on edge devices (Oviatt, 1999).

ii. Ethical and Social Concerns: The collection of visual and auditory data raises significant privacy and security issues (Khan Mohd et al., 2022). Furthermore, models can perpetuate or amplify societal biases present in training data, leading to unfair performance across demographic groups.

iii. Real-World Deployment: Challenges include hardware limitations, user acceptance of always-on sensors, and maintaining robustness in diverse, unstructured environments.

Future Directions

i. Adaptive and Context-Aware Fusion: Developing systems that dynamically prioritize modalities based on environmental context (e.g., favoring gestures in a loud room) or task requirements.

ii. Cross-Cultural and Inclusive Design: Systematically addressing cultural variability in gestures, expressions, and communication norms to ensure global usability (Cassell, 2001).

iii. Ethical and Responsible AI: Advancing frameworks for transparent consent, data anonymization, algorithmic fairness audits, and bias mitigation throughout the AI lifecycle.

iv. Integration of Novel Modalities: Exploring the additive value of haptic, physiological, and even olfactory data to create ultra-personalized and immersive interaction loops.

v.

Conclusion

Multimodal conversational AI represents a significant leap toward bridging the gap between human and machine communication. By integrating text, voice, and visual data, and potentially beyond, these systems enable richer, more context-aware, and empathetic interactions. Frameworks like SIMMC and models like DialogueTRM demonstrate substantial progress in task-oriented and affective computing domains. However, persistent challenges in data synchronization, computational scaling, privacy, and bias must be rigorously addressed to achieve trustworthy and widespread adoption. Future progress hinges on developing more adaptive, ethical, and culturally-aware systems, supported by interdisciplinary collaboration across AI, HCI, ethics, and social sciences.

References

Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(2), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607

Cassell, J. (2001). Embodied conversational agents: Representation and intelligence in user interfaces. *AI Magazine, 22*(4), 67–83. https://doi.org/10.1609/aimag.v22i4.1593

Feldman, S., Yalcin, O. N., & DiPaola, S. (2017). Engagement with artificial intelligence through natural interaction models. In *Electronic Visualisation and the Arts (EVA 2017)* (pp. 296–303). BCS Learning & Development Ltd. https://doi.org/10.14236/ewic/EVA2017.60

Khan Mohd, T., Nguyen, N., & Javaid, A. Y. (2022). Multi-modal data fusion in enhancing human-machine interaction for robotic applications: A survey. *arXiv*. https://arxiv.org/abs/2202.07732

Korbar, B., Tran, D., & Torresani, L. (2018). Cooperative learning of audio and video models from self-supervised synchronization. *Advances in Neural Information Processing Systems, 31*, 7763–7774. https://papers.nips.cc/paper_files/paper/2018/hash/c4616f5a24a66668f11ca4fa80525dc4-Abstract.html

Liang, P. P., Zadeh, A., & Morency, L.-P. (2024). Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys, 56*(10), Article 264, 1–42. https://doi.org/10.1145/3656580

Mao, Y., Sun, Q., Liu, G., Wang, X., Gao, W., Li, X., & Shen, J. (2020). DialogueTRM: Exploring the intra- and inter-modal emotional behaviors in the conversation. *arXiv*. https://arxiv.org/abs/2010.07637

McTear, M. F. (2017). The rise of the conversational interface: A new kid on the block? In J. F. Quesada, F. J. Martín Mateos, & T. López-Soto (Eds.), *Future and emerging trends in language technology: Machine learning and big data* (pp. 38–49). Springer. https://doi.org/10.1007/978-3-319-69365-1_3

Moon, S., Kottur, S., Crook, P., De, A., Poddar, S., Levin, T., Whitney, D., Difranco, D., Beirami, A., Cho, E., Subba, R., & Geramifard, A. (2020). Situated and interactive multimodal conversations. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 1103–1121). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.96

Oviatt, S. (1999). Ten myths of multimodal interaction. *Communications of the ACM, 42*(11), 74–81. https://doi.org/10.1145/319382.319398

Potamianos, G., Neti, C., Gravier, G., Garg, A., & Senior, A. W. (2003). Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE, 91*(9), 1306–1326. https://doi.org/10.1109/JPROC.2003.817150

Schwartz, I., Yu, S., Hazan, T., & Schwing, A. G. (2019). Factor graph attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2039–2048). IEEE. https://doi.org/10.1109/CVPR.2019.00214