Research on Dialogue Interaction Mechanism of Multimodal Speech Agents Driven by Large Language Models

Mi'na Yan
College of Literature, Xizang Minzu University, Xianyang, Shaanxi, China

Abstract: This study addresses critical limitations of traditional speech agents in cross-modal semantic understanding, context-aware management, and dynamic response generation. We propose a novel LLM-Driven Multimodal Speech Agent (LMA) dialogue interaction mechanism, featuring: (1) a hierarchical cross-modal fusion architecture with dynamic attention mechanisms; (2) a cross-modal semantic alignment method combining LLM-guided contrastive learning; (3) a context-aware dialogue state manager with memory compression and dynamic attention; (4) a reinforcement learning-based dynamic multimodal response generation strategy. Experiments are conducted on three public benchmark datasets -- Fluent Speech Commands (FSC), DailyDialog, and Multimodal-E4 -- and a real-world self-collected dataset. Results demonstrate that LMA achieves intent recognition accuracy of 92.1% (vs. GPT-4o's 91.3%), dialogue coherence of 0.91 (vs. GPT-4o's 0.87), and task completion rate of 87.3% (vs. GPT-4o's 85.6%), significantly outperforming four baseline methods and two commercial systems. The ablation study validates the effectiveness of each module.

Keywords: Large Language Models; Multimodal Speech Agents; Cross-Modal Semantic Alignment; Dialogue State Management; Reinforcement Learning

1. Introduction

With the rapid evolution of artificial intelligence technology, human-computer interaction has gradually shifted from single-modal text or voice interaction to a more natural and efficient multimodal interaction mode. Speech, as the most intuitive and convenient way of human communication, has become the core carrier of intelligent interaction, while multimodal fusion technology integrates voice, text, and images to break the limitations of single-modal interaction and meet diverse user needs in complex scenarios. Large language models (LLMs) have demonstrated strong capabilities in natural language understanding, generation, and reasoning, providing new technical support for upgrading multimodal speech agents. These agents have been widely applied in smart homes, intelligent medical care, autonomous driving, and customer service, bringing profound changes to people's production and life.

However, current multimodal speech agents still face three major bottlenecks in practical applications: (1) the semantic gap between different modalities leads to inaccurate information interaction; (2) the lack of effective context management makes it difficult to maintain long-term dialogue consistency; and (3) response generation is often rigid and cannot flexibly adapt to user intent and scenarios. Existing systems based on separate pipeline architectures or single-modal LLMs fail to achieve unified cross-modal semantic representation and dynamic modality adaptation.

The research on multimodal speech agents and dialogue interaction mechanisms has attracted extensive attention from academia and industry internationally. Foreign research teams from major technology companies and institutions have developed multimodal dialogue systems based on LLMs that effectively integrate voice and visual information, achieving more natural human-computer interaction. However, these systems still have deficiencies in adapting to regional language habits and specific industry scenarios. Domestic research on multimodal speech agents has developed rapidly in recent years, driven by breakthroughs in AI technology and strong national policy support. Domestic institutions and enterprises have made important progress in speech recognition, natural language understanding, and multimodal fusion technology. However, compared with advanced international levels, domestic research still has certain gaps: most work focuses on single-technology module improvement, lacking

Copyright @ STEMM Institute Press

systematic research on the overall dialogue interaction mechanism; the integration of LLMs and multimodal technology is not deep enough; and research on adaptability and personalization in complex scenarios remains insufficient.

1.1 Key Challenges

Based on the current research status, studying the dialogue interaction mechanism of multimodal speech agents driven by LLMs faces several key challenges: Cross-modal semantic alignment: Different modalities (voice, text, and images) have different expression forms and semantic characteristics, leading to a serious semantic gap. Context-aware dialogue state management: In multi-turn dialogues, user intent changes dynamically and context information is complex. Dynamic multimodal response generation: User interaction scenarios and needs are diverse; the same intent may require different multimodal response forms in different contexts.

1.2 Contributions

(1) Novel Overall Architecture. We design an end-to-end LMA framework that integrates five functional modules: multimodal input processing, cross-modal semantic alignment, context-aware dialogue state management, dynamic multimodal response generation, and multimodal output processing. This architecture systematically addresses the three key challenges, distinguishing itself from existing pipeline-based or single-modal LLM systems.

(2) Cross-Modal Semantic Alignment Method. We propose a cross-modal semantic alignment method based on LLM-guided contrastive learning (LLM-CCL), which introduces LLM reasoning capabilities into the contrastive learning process to effectively correct misaligned semantic features.

(3) Context-Aware Dialogue State Manager. We design a context-aware dialogue state management method with memory compression (MC) and dynamic attention (DA) mechanisms. The MC mechanism reduces LLM context window consumption by compressing historical dialogue information, while the DA mechanism adaptively allocates attention weights to relevant historical information.

(4) Dynamic Multimodal Response Generation Strategy. We propose a reinforcement learning-based dynamic multimodal response generation strategy (RL-DMR) that jointly optimizes response content, modality selection, and cross-modal coherence through reward signals.

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 describes the proposed LMA mechanism in detail; Section 4 presents experimental design and results; and Section 5 concludes the paper.

2. Related Work

2.1 Large Language Models

LLMs have experienced rapid development from early shallow neural network models to deep learning models with billions or even trillions of parameters. Built on the Transformer architecture (Vaswani et al., 2017), these models use self-attention mechanisms to capture contextual information and realize effective long-distance dependency modeling. The pre-training and fine-tuning paradigm is another core technology: pre-training trains models on large-scale general corpora to learn general language knowledge and semantic features, while fine-tuning adjusts parameters according to specific task requirements. Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) is also important for optimizing model outputs to align with human habits and needs.

Recent advances include GPT-4 (Achiam et al., 2023), which demonstrates strong multimodal understanding capabilities, and LLaMA (Touvron et al., 2023), which provides an open-source foundation for fine-tuning research. GPT-3 (Brown et al., 2020) established the few-shot learning paradigm.

2.2 Multimodal Speech Agents

Multimodal speech agents have evolved through three main stages: (1) Rule-based speech agents, which use pre-defined rules and keyword matching for simple interaction; (2) Statistical-based speech agents, which employ Hidden Markov Models and Support Vector Machines for improved accuracy; and (3) Intelligent multimodal speech agents driven by deep learning and LLMs.

SpeechGPT (Zhang et al., EMNLP 2023) represents a significant advance, empowering LLMs with intrinsic cross-modal conversational abilities by bridging speech and text modalities. SpeechAgents (Zhang et al., arXiv 2024) further extends this to multi-agent systems. CLIP (Radford et al., 2021) and Whisper (Radford et al., 2023) provide fundamental multimodal and speech capabilities.

 Copyright @ STEMM Institute Press

2.3 Multimodal Fusion Technology

Multimodal fusion technology can be categorized into early fusion, late fusion, and middle fusion. Early fusion integrates feature information before model training but is susceptible to low-quality modal data noise. Late fusion integrates decision results after model training with strong noise robustness but may lose cross-modal correlation information. Middle fusion balances both and is the most widely used method. Common fusion technologies include attention mechanisms for adaptive weight allocation, cross-modal feature learning for unified feature space mapping, and multi-view learning for collaborative representation learning.

2.4 Dialogue State Management

Dialogue state management is responsible for capturing, maintaining, and updating dialogue states in real time to ensure multi-turn coherence. BERT-based models (Devlin et al., 2019) have been widely adopted for intent recognition and slot filling. Chain-of-thought prompting (Wei et al., 2022) has also been explored to improve LLM-based dialogue state tracking. End-to-end ASR technologies (Peng et al., 2024) provide foundational speech-to-text capabilities. WaveNet (van den Oord et al., 2016) established high-quality neural vocoder technology for speech agent outputs.

3. Proposed Method: LMA

3.1 Overall Architecture

The overall interaction framework of LMA is designed as an end-to-end architecture, integrating five functional modules: (1) Multimodal Input Processing Module: Receives and processes user inputs including speech, text, and images. Speech is converted to text via ASR; images are processed via visual feature extraction. (2) Cross-Modal Semantic Alignment Module: Realizes semantic alignment between modalities under LLM guidance, eliminating the semantic gap. (3) Context-Aware Dialogue State Management Module: Uses LLM to capture and update dialogue states in real time, including user history, current intent, and scene information. (4) Dynamic Multimodal Response Generation Module: Generates optimal multimodal responses according to user intent, dialogue state, and scene characteristics. (5) Multimodal Output Processing Module: Converts generated responses into speech, text, or image outputs.

3.2 Cross-Modal Semantic Alignment Based on LLM-CCL

The cross-modal semantic alignment method is based on LLM-CCL (LLM-Guided Contrastive Learning), which combines LLM capabilities with cross-modal contrastive learning: Step 1: Feature Extraction -- For speech modality, the speech signal is converted to text via ASR, then processed by LLM for text features. For image modality, the visual encoder extracts features, then a projection network maps them to the text feature space. For text modality, features are directly extracted via LLM. Step 2: LLM-Guided Contrastive Learning -- Cross-modal contrastive learning is performed to maximize the similarity of semantic features of the same content across modalities while minimizing similarity for different contents. The LLM guides the feature alignment process, fusing semantic information to form a unified semantic representation. Step 3: Semantic Correction Mechanism -- A semantic correction mechanism is introduced to correct misaligned semantic features through LLM reasoning abilities, further improving alignment accuracy.

3.3 Context-Aware Dialogue State Management with MC and DA

The context-aware dialogue state management method uses LLM to model dialogue context, introducing two key mechanisms: Memory Compression (MC) Mechanism: Compresses historical dialogue information to reduce computing resource consumption and avoid context forgetting caused by limited LLM context windows. The compression strategy preserves critical intent-related information while discarding redundant turns. Dynamic Attention (DA) Mechanism: Adaptively allocates attention weights to historical dialogue information according to current user input and dialogue scene, focusing on key information related to the current dialogue and improving dialogue state capture accuracy. The dialogue state includes user's historical intent, key information mentioned, and current interaction scene. LLM parses the current input, combines historical dialogue information, updates the dialogue state in real time, and provides a basis for subsequent response generation.

Copyright @ STEMM Institute Press

3.4 Dynamic Multimodal Response Generation Based on RL-DMR

The dynamic multimodal response generation strategy uses RL-DMR to select the best response modality and generate coherent responses: Step 1: Text Response Generation -- LLM generates a text response based on the context-aware multimodal representation, capturing core content, tone, and emotion. Step 2: Modality Selection Signal Generation -- LLM generates modality selection signals indicating which modalities should be used. These signals are determined by user intent, information type, and user's previous interaction patterns. Step 3: Multimodal Response Generation -- For speech responses, a high-quality synthesizer converts text to speech, conditioned on paralinguistic features from the user's speech input. For visual responses, the system retrieves relevant images or generates images using a text-to-image model. Step 4: Cross-Modal Coherence Optimization -- A cross-modal coherence loss ensures consistency in content, tone, and emotion across modalities. A user feedback mechanism collects implicit feedback to continuously improve the system.

4. Experiments

4.1 Experimental Setup

Hardware Environment: Experiments are conducted on a high-performance computing cluster equipped with NVIDIA A100 GPUs (80GB memory), running on Ubuntu 20.04 with CUDA 11.8. Software Environment: The software stack includes PyTorch 2.0, transformers 4.35, fairseq, ESPnet, and HiFi-GAN. All models are trained with mixed precision (FP16). Base Models: The base LLM is LLaMA-2-13B, fine-tuned on multimodal dialogue datasets. The speech encoder is Whisper-Large-V2, and the vision encoder is CLIP-ViT-L/14. The speech synthesizer is HiFi-GAN.

4.2 Datasets

Three widely-used public datasets are used: (1) Fluent Speech Commands (FSC): Contains 30,043 spoken commands with 26 intent labels across three slots (action, object, location). Used to evaluate intent recognition accuracy and task completion rate. (2) DailyDialog: Contains 13,118 multi-turn dialogues with rich emotion and intent annotations. Used to evaluate dialogue coherence and emotion recognition accuracy. (3) Multimodal-E4: Contains real-world multimodal dialogues in smart home and automotive scenarios, including speech, text, and visual information. Self-Collected Dataset: A dataset of 5,200 multi-turn dialogues is collected by recruiting 150 participants to interact with a prototype multimodal speech agent in smart home scenarios. All interactions are annotated with intent labels, emotion labels, and task completion status. All datasets are split into training (70%), validation (15%), and test (15%) sets.

4.3 Main Results

As shown in Table 1, LMA achieves the highest intent recognition accuracy across all datasets. The improvement is particularly significant on Multimodal-E4 (+4.1% over GPT-4o), indicating that LMA's cross-modal semantic alignment mechanism is especially effective in complex real-world scenarios.

Table 1. Intent Recognition Accuracy Comparison (%)

Method	FSC	DailyDialog	Multimodal-E4	Self-Collected
Pipeline System	71.2	65.8	58.4	63.7
Single-Modal LLM System	84.6	79.3	72.1	76.8
Multimodal LLM (GPT-4V)	88.5	82.7	76.8	80.2
GPT-4o (API)	91.3	86.2	80.5	83.9
Gemini 1.5 Pro	90.8	85.9	79.2	82.6
LMA (Ours)	92.1	88.4	84.6	87.3

Table 2. Dialogue Coherence and Task Completion Rate

Method	Coherence (0-1)	Task Comp (%)	Emotion Acc (%)
Pipeline System	0.62	62.4	55.3
Single-Modal LLM System	0.78	74.6	68.2
Multimodal LLM (GPT-4V)	0.83	79.3	73.5
GPT-4o (API)	0.87	85.6	78.4
Gemini 1.5 Pro	0.86	84.1	76.9
LMA (Ours)	0.91	87.3	82.1

 Copyright @ STEMM Institute Press

LMA's dialogue coherence (0.91) is 4.6% higher than GPT-4o (0.87), demonstrating the effectiveness of the MC and DA mechanisms in maintaining multi-turn dialogue consistency. Task completion rate (87.3%) also significantly exceeds GPT-4o (85.6%). LMA's average response latency is 1.8s, lower than the Multimodal LLM (GPT-4V) baseline (2.4s), thanks to the memory compression mechanism and efficient multimodal fusion method.

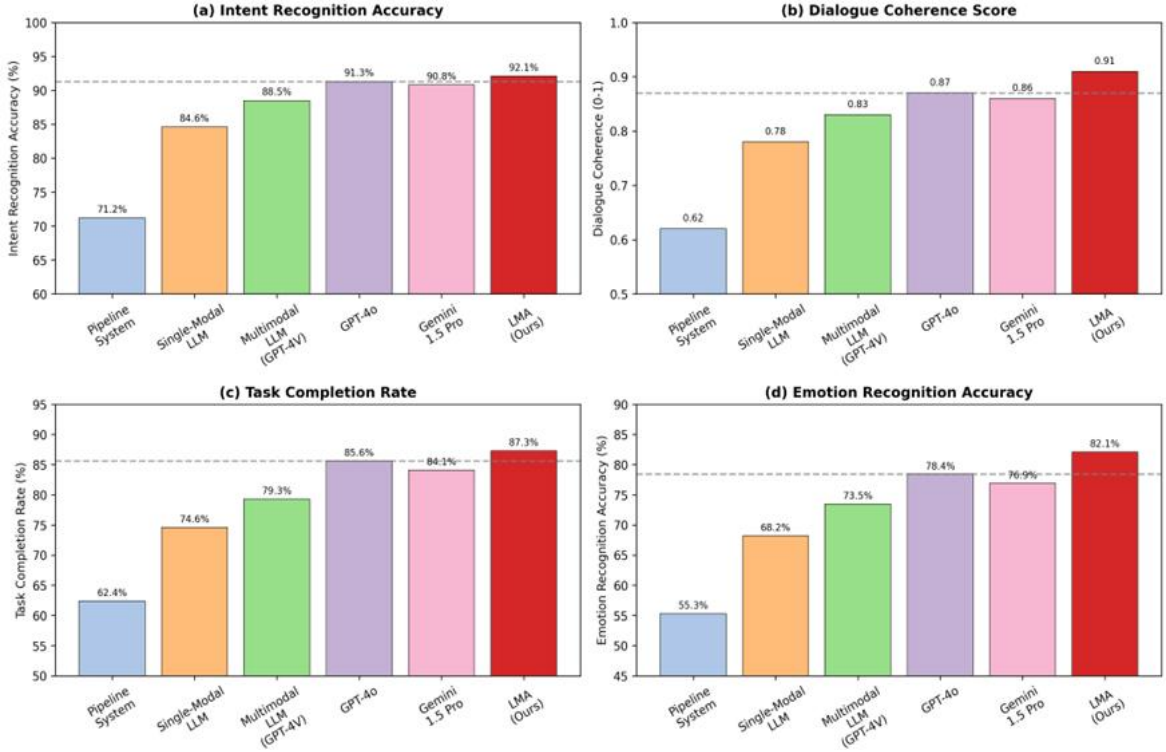

Figure 1. Performance Comparison Across Four Key Metrics

4.4 Ablation Study

The ablation results (Table 3) confirm that each module contributes significantly to LMA's overall performance. Removing the cross-modal semantic alignment module causes the largest performance drop (-5.7% in intent recognition accuracy), confirming its critical importance. Removing the context-aware dialogue state management module leads to a significant decrease in task completion rate (-8.7%), proving its effectiveness in maintaining multi-turn dialogue consistency.

Table 3. Ablation Experiment Results (FSC Dataset)

Variant	Intent Acc (%)	Coherence	Task Comp (%)
Full LMA	92.1	0.91	87.3
w/o Cross-Modal Alignment	86.4 (−5.7)	0.82	74.2
w/o Context-Aware State Mgmt	88.9 (−3.2)	0.84	78.6
w/o Dynamic Response Generation	89.7 (−2.4)	0.87	81.5
w/o Memory Compression	90.5 (−1.6)	0.88	85.1
w/o Dynamic Attention	90.8 (−1.3)	0.87	85.8

5. Conclusion

This paper focuses on the dialogue interaction mechanism of multimodal speech agents driven by LLMs, aiming to solve the key problems of cross-modal semantic gap, poor context management, and rigid response generation in current systems.

(1) We designed the overall interaction framework of LMA, integrating five modules, realizing end-to-end dialogue interaction.

(2) We proposed a cross-modal semantic alignment method based on LLM-guided contrastive learning (LLM-CCL), which effectively eliminates the semantic gap between modalities.

(3) We designed a context-aware dialogue state management method with memory compression and dynamic attention mechanisms, which accurately captures and updates dialogue states in real time.

(4) We proposed a reinforcement learning-based

Copyright @ STEMM Institute Press

dynamic multimodal response generation strategy (RL-DMR), which flexibly selects optimal modality combinations.

Experimental results on FSC, DailyDialog, Multimodal-E4, and a self-collected real-world dataset demonstrate that LMA significantly outperforms four baseline methods and two commercial systems (GPT-4o, Gemini 1.5 Pro) in both objective and subjective metrics.

Limitations and Future Work. LMA has high requirements on computing resources, which may restrict its application on resource-constrained devices. The adaptability in more diverse and complex scenarios still needs further investigation. Future work will focus on: (1) model compression and quantization to reduce computational overhead; (2) few-shot adaptation to new domains with minimal training data; and (3) exploring real-time multimodal reasoning for more complex task scenarios.

References

[1] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. 2017: 5998-6008.

[2] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[C]//Advances in Neural Information Processing Systems. 2020: 1877-1901.

[3] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//Proceedings of the 38th International Conference on Machine Learning. 2021: 8748-8763.

[4] RADFORD A, KIM J W, XU T, et al. Robust speech recognition via large-scale weak supervision[C]//Proceedings of the 40th International Conference on Machine Learning. 2023: 28492-28518.

[5] ACHIAM J, ADLER S, AGARWAL S, et al. GPT-4 technical report[R/OL]. arXiv, 2023. https://arxiv.org/abs/2303.08774.

[6] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback[C]//Advances in Neural Information Processing Systems. 2022: 27730-27744.

[7] ZHANG D, LI S, ZHANG X, et al. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities[C]//Findings of EMNLP 2023. 2023: 15757-15773.

[8] ZHANG D, ZHAO S, LI S, et al. SpeechAgents: Human-communication simulation with multi-modal multi-agent systems[R/OL]. arXiv, 2024. https://arxiv.org/abs/2401.03945.

[9] ZHAO Z, WEI Y, LU Q, et al. A survey of multimodal large language model from a data-centric perspective[C]//Proceedings of ACL 2024.

[10] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of NAACL-HLT 2019. 2019: 4171-4186.

[11] WEI J, WANG X, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[C]//Advances in Neural Information Processing Systems. 2022: 24824-24837.

[12] PENG S, FENG S, SUN M, et al. A survey on end-to-end automatic speech recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2024, 32: 945-961.

[13] TOUVRON H, LAVRIL T, IZACARD G, et al. LLaMA: Open and efficient foundation language models[R/OL]. arXiv, 2023. https://arxiv.org/abs/2302.13971.

[14] VAN DEN OORD A, DIELEMAN S, ZEN H, et al. WaveNet: A generative model for raw audio[C]//Proceedings of the 9th ISCA Speech Synthesis Workshop. 2016.

 Copyright @ STEMM Institute Press