# Improving Language Learning Performance Using Multimodal Dialogue Systems

Zhenyu Wu[1](✉), Zhiyang Ding[1], Zhaowei Zhang[1], and Yanqin Mao[2]

[1] School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing, China
zhenyu.wu@njupt.edu.cn

[2] School of Software, Nanjing University of Posts and Telecommunications, Nanjing, China

**Abstract.** Learning a language poses a significant challenge for learners, who must comprehend the intricacies of language and analyze the relationships among its components. While computer technologies have been developed to assist language learning, they fail to mirror the human cognitive process. This paper examines the application of a multimodal dialogue system to enhance language learning outcomes. The system boasts several advantages. Firstly, smart devices can collect multimodal data in learning environments to monitor the learner's status in real-time, thus enhancing the accuracy of intention recognition. Secondly, the system can interact with learners naturally by analyzing their multimodal data, resulting in improved language skills. Finally, application scenarios are designed based on the defined multimodal dialogue system, which effectively demonstrates the system's ability to enhance language learning performance.

**Keywords:** language learning · natural language processing · multimodal interaction · multimodal dialogue system

## 1 Introduction

Computers have been used for language learning and teaching since the 1960s, thanks to their powerful calculation and storage capabilities [1]. In the last 20 years, with the development of information technologies such as the Internet, mobile applications, big data, and artificial intelligence, computer-assisted language learning (CALL) has become increasingly popular and important. Virtual reality, digital games, and virtual agents are a few examples of the ways technology has been applied to language learning. However, the most challenging and significant development in CALL is natural language processing (NLP), which aims to formalize natural languages using mathematical models. Since professor Manning proposed computational linguistics [5], NLP technology has made significant progress. The latest models, such as ChatGPT [6], have achieved excellent results in human-machine question-answer conversations, leading us to ponder the relationship between technology and language learning.

To enhance language learning through the use of these artificial intelligence models, we must consider two-way interactions between human learners and the models. First,

F. Huang et al. (Eds.): ICAIE 2023, AHCS 15, pp. 191–197, 2024.
https://doi.org/10.2991/978-94-6463-242-2_23

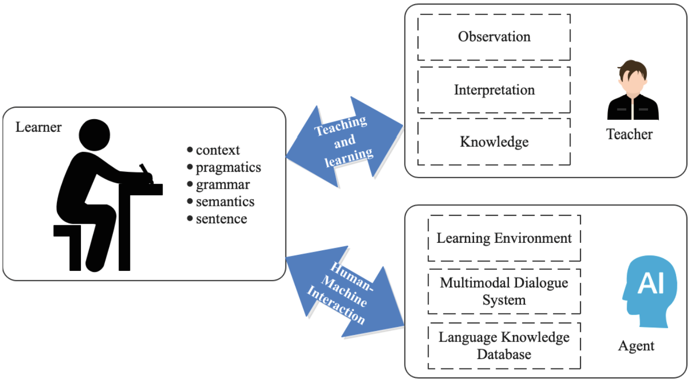

**Fig. 1.** Learning language from teachers and artificial intelligence agents

the models should learn about the status and behavior of learners in order to provide more suitable services. Second, personalized learning is key. A good CALL system should recommend personalized learning content based on a learner's personal profile, learning history, and learning status. This approach reduces the cognitive load on learners and improves their learning performance. Finally, it is essential to maintain the language context, as learning is a social process [8]. For instance, a conversation between a teacher and a student learning a language in a classroom context is shown in Fig. 1.

## 2 Related Works

At the early stage of CALL, computers were primarily used for repetitive language drills such as grammatical explanations and translation tests [9]. As learning theories evolved, students were encouraged to generate original utterances, leading to the emergence of communicative CALL, which focused on the cognitive process of learning [10]. However, at that time, computers weren't playing adequate roles in the language learning process. Teachers emphasized that language should be used in authentic social contexts and that various language learning skills should be integrated. This gave rise to a new field: integrative CALL [11]. Thanks to the advancement of the Internet and mobile communication, computer technologies have greatly facilitated the development of language learning. For instance, Liu et al. developed a mobile application to help English learners improve their listening comprehension [12].

The use of dialogue systems is becoming increasingly popular, and researchers are starting to pay more attention to multimodal dialogue systems. In 1999, Kuo et al. defined a multimodal dialogue system that could accept input modalities such as text, images, and voice [13]. This allowed users to communicate with the system using both language and nonverbal signals, such as gestures. Kuansan et al. proposed a multimodal dialogue framework that used the inputted multimodal interaction data to infer dialogue intentions of users [14]. Zhou et al. proposed a multimodal dialogue framework that included video conference capabilities [15]. Although there have been several studies on multimodal dialogue systems, their incorporation into language learning has not been widely reported.

## 3 Multimodal Dialogue System for Language Learning

(1) Learning environment

The learning environment comprises the learner and smart devices such as microphones and cameras. By continuously interacting with a multimodal dialogue system, learners can improve their language capabilities. This framework is illustrated in Fig. 2.

The microphone captures the voice signals of the learner, while the camera records their learning behaviors, including hand gestures, body gestures, and facial expressions. This recorded information is digitalized, saved, and delivered to the multimodal dialogue system for further analysis.

(2) Multimodal dialogue system

The inputted voice signal is first processed by an automatic speech recognition (ASR) module, where the content and emotional status are recognized by machine learning models. Similarly, the inputted camera data is processed by a vision analysis module, where hand gestures, body gestures, and facial expressions are recognized using the latest deep learning models, such as convolutional neural networks (CNNs). Moreover, learners may directly input text to communicate with the multimodal dialogue system. These three forms of multimodal data reflect the real learning requirements of learners.

To identify the specific intentions of learners, a user intention recognition module is used to process the results returned by the ASR, vision analysis, and text input. Here, multimodal machine learning could be applied to complete this task. The intentions may include "asking questions," "querying grammar," "searching for words," and "practicing."

After identifying the intentions of learners, the multimodal dialogue management model determines how to respond using historical data collected from past interactions. This model maintains the dialogue status and predicts future dialogue actions using tools such as the hidden Markov model (HMM) and reinforcement learning model. However, training these machine learning models using collected historical multimodal data is a significant challenge.

The database module holds a large amount of language data, which can be used for generating responses. It should be a multimodal database capable of storing multimodal

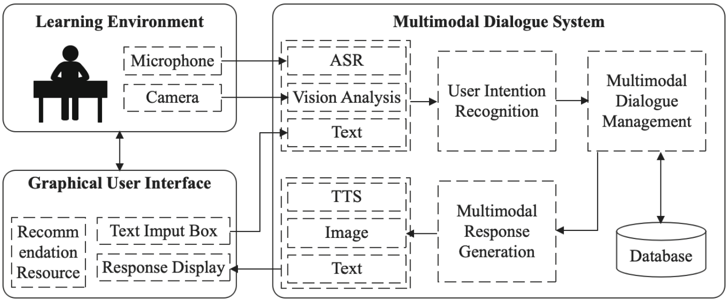

**Fig. 2.** Framework of multimodal dialogue system for language learning

data such as texts, images, voices, and videos. Additionally, the database can take the form of a multimodal knowledge graph, which has a more robust capability of organizing multimodal data and their relationships.

The multimodal generation module is responsible for producing the final response. Compared to traditional natural language generation technologies, the results of the multimodal generation module may include images, videos, and texts. The latest generative artificial intelligence technologies can be used in this module to achieve better results.

The generated texts can be directly returned to learners through a graphical user interface, or they can be transformed into voice signals using text-to-speech (TTS) technologies. The voice signals can be played through the stereo to provide audio feedback to learners. Furthermore, the generated images can be displayed in the graphical user interface. Both the voice signals and images can more vividly meet the needs of learners.

(3) Graphical user interface

The graphical user interface can display the operations of learners, with a text input box for accepting text input from learners. Responses generated by the multimodal dialogue system, such as images, texts, and other media, can be displayed in the response display area.

# 4 Application Scenarios

## 4.1 Language Learning via Dialogues

Language learning through dialogues is illustrated in Fig. 3. The camera and microphone are used to capture the status of the learner, who communicates with the agent via text and voice. The agent returns responses to the learner, facilitating language learning for both parties. Continuous interaction like this is expected to improve the language skills of learners.

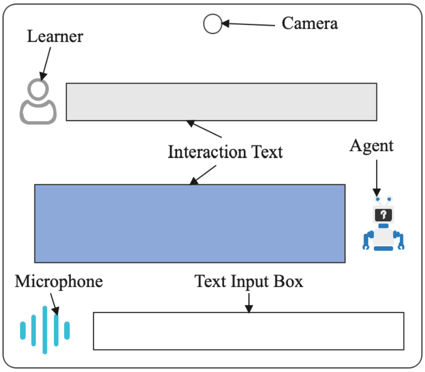

**Fig. 3.** Language learning via dialogues

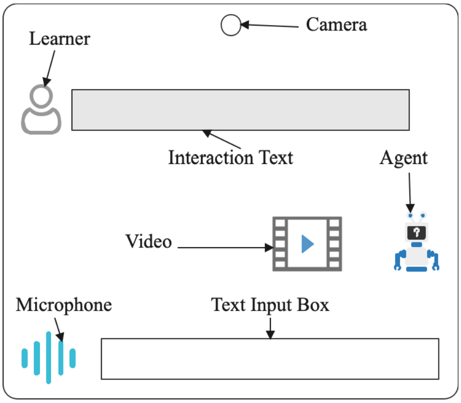

**Fig. 4.** Language learning via multimodal data

### 4.2 Language Learning via Multimodal Data

Typically, language is closely associated with textual data, including words, sentences, paragraphs, and articles. Thus, texts are crucial for conveying information among human beings. However, in the case of language learning, other types of modal data can also aid in improving learning performance.

In Fig. 4, the agent can generate a video to respond to the learner's query. This video can be a digital resource that explains concepts related to the learner's questions. By watching the video, the learner can gain a deeper understanding of the concepts. In addition to videos, the agent can also provide images and voice signals as responses. The modality and content of the response data can be determined based on the user profiles and learning status of the learners. For instance, if the agent analyzes the captured facial expression data and determines that the learner is tired, it may play a video to grab the learner's attention. Additionally, the agent can ask the learner to complete a gesture by giving a language command. By accurately listening and understanding the language command, the learner can perform the gesture correctly, which can further promote enthusiasm for learning.

## 5 Conclusion

In this paper, we have explored the language learning process and introduced a multimodal dialogue system that can be applied in language learning. This system employs smart devices such as cameras and microphones to collect multimodal data about the learner's status. This data is invaluable for identifying the learner's intentions. Moreover, multimodal dialogue management can generate feasible responses by interacting with language knowledge databases. Finally, learners can receive their preferred responses in various forms, such as texts, images, and videos. The graphical user interface facilitates the relationship between the learners and the multimodal dialogue system.

We also introduced two application scenarios where multimodal dialogue systems can be effectively used for language learning: language learning through dialogues and language learning using multimodal data. By using these systems, learners can improve their language skills and achieve their learning goals more effectively.

**Acknowledgements.** This work was supported by National Natural Science Foundation of China (No. 61502246), and Nanjing University of Posts and Telecommunications Science Foundation (Nos. NY215019, NY220202).

## References

1. M. Warschauer and D. Healey, "Computers and language learning: an overview," Language teaching, vol. 31, pp. 57-71, April 1998.
2. P. Li, J. Legault, A. Klippel, and J. Zhao, "Virtual reality for student learning: Understanding individual differences," Human Behaviour and Brain, vol. 1, pp. 28-36, March 2020.
3. E.O. Acquah and H.T. Katz, "Digital game-based L2 learning outcomes for primary through high-school students: A systematic literature review," Computers & Education, vol. 143, pp. 103667, January 2020.
4. J. Junaidi, B. Hamuddin, K. Julita, F. Rahman, and T. Derin, "Artificial intelligence in EFL context: Rising students' speaking performance with Lyra Virtual Assistance," International Journal of Advanced Science and Technology Rehabilitation, vol. 29, pp. 6735-6741, 2020.
5. C. Manning and H. Schutze, Foundations of Statistical Natural Language Processing. MIT Press, 1999.
6. N. Stiennon, L. Ouyang, W. Jeffrey, et al, "Learning to summarize with human feedback," Advances in Neural Information Processing Systems, vol. 33, pp. 3008-3021, 2020.
7. R. Krishna, D. Lee, L. Fei-Fei, and M.S. Bernstein, "Socially situated artificial intelligence enables learning from human interaction," Proceedings of the National Academy of Sciences, vol. 119, p. e2115730119, September 2022.
8. B. Norton and K. Toohey, "Identity, language learning, and social change," Language Teaching, vol. 44, pp. 412-446, October 2011.
9. K. Ahmad, G. Corbett, M. Rogers, and R. Sussex, Computers, language learning and language teaching. Cambridge: Cambridge University Press, 1985, pp. vii+ 158.
10. C. Jones and S. Fortescue, Using computers in the language classroom. London: Longman, 1987.
11. M. Warschauer, "Computer-assisted language learning: an introduction," Multimedia language teaching, vol. 320, March 1996.
12. G.Z. Liu, J.Y. Chen, and G.J. Hwang, "Mobile-based collaborative learning in the fitness center: A case study on the development of English listening comprehension with a context-aware application," British Journal of Educational Technology, vol. 49, pp. 305-320, March 2018.
13. A. Potamianos, H.K. Kuo, C.H. Lee, et al, "Design Principles and Tools for Multimodal Dialog Systems," In ESCA Tutorial and Research Workshop (ETRW) on Interactive Dialogue in Multi-Modal Systems, 1999.
14. K. Wang, "Implementation of a Multimodal Dialog System Using Extended Markup Languages," In Sixth International Conference on Spoken Language Processing, 2000.
15. Z. Yu, V. Ramanarayanan, R. Mundkowsky, et al, "Multimodal HALEF: An Open-Source Modular Web-Based Multimodal Dialog Framework," Dialogues with Social Robots: Enablements, Analyses, and Evaluation, pp. 233-244, 2017.