DOI: https://doi.org/10.3217/978-3-85125-609-3-13

Building Human-Scale Intelligent Immersive Spaces for Foreign Language Learning

Rahul R. Divekar(✉)[1], Yalun Zhou[1], David Allen[1], Jaimie Drozdal[1], and Hui Su[1,2]

[1] Rensselaer Polytechnic Institute, Troy, New York, USA
[2] IBM T.J. Watson Research Center, Yorktown Heights, New York, USA
{divekr, zhouy12, allend5, drozdj3}@rpi.edu, huisuibmres@us.ibm.com

Abstract. Immersive technologies (e.g., human-scale immersion) especially augmented with Artificial Intelligence (AI) have been sparingly used to teach foreign languages. This work discusses an AI equipped immersive room to teach Mandarin Chinese and, by doing so it brings two perspectives together - the pedagogy of foreign language teaching and the technology that has been built to support its goals. We therefore have translated specific teaching outcomes and requirements of foreign language learning into technology problems and brought forth a Cognitive Immersive Room (CIR) - a human-scale immersive room equipped with an AI Agent to teach foreign language. We have discussed a pilot study with 16 college student participants to show its usability and explore ways of how the room could be used as a part of the foreign language curriculum.

Keywords: Immersive Learning · Chinese as a foreign language · AI assisted learning.

1 Introduction

The use of immersion in teaching foreign languages is highly praised. We define virtual immersion for our context as visual, cultural and linguistic. It is enabled with multi-modal interaction in its input and output (e.g. Speech and gesture for input. Audio-visuals for output). One can imagine the use of a context-rich environment, giving users the feeling of almost being in a foreign country, as they practice their foreign language in a natural way and complete meaningful tasks.

In this work, we elaborate on a 360 degree human-scale panoramic display that is equipped with a virtual AI agent who can "see", "hear" and "talk" to its occupants (students in this case). We have aimed to build an experience that is as close as possible to traveling to a foreign country and allowing students to practice their spoken conversational skills. However, building the experience of living in a foreign country is not enough for this use-case until the AI Agent can actually aid language learning. We thus describe the functions in this room, and how they specifically target requirements and outcomes defined by task-based

DOI: https://doi.org/10.3217/978-3-85125-609-3-13

language learning pedagogy, yet keep the experience natural. This work therefore presents a multi-disciplinary approach between Artificial Intelligence, Immersion and Foreign Language Pedagogy where pedagogy has driven technology design and development.

In the next sections, we review the existing literature in the domain, briefly describe the technology we have built to give readers a broad context, outline the learning expectations from the pedagogical aspects and describe the specific features of the system that were driven by pedagogical expectations. We show in a pilot-study that the immersive environment was likable and usable, and shed light on the future directions to make this a successful learning experience.

2 Literature Review

Virtual Immersive learning has been hailed by several experts in the community. Most of the discussion in this domain revolves around VR/AR applications. In this section, we only briefly touch upon the trends in this space by meta-reviewing review papers. This is not an exhaustive list of applications out there. Such a list can be found in papers specifically meant to review the space of immersion and technology, examples of which are [1] and [2].

We see two trends in the discussions; using multi-media (e.g videos) as immersion and using AR/VR applications to achieve immersion - both possibly aimed at presenting knowledge and practice to the students in a multi-modal way (e.g sound, graphics, etc.). We realize the importance of context-richness as [3] have shown that such a multi-modal presentation is effective in second language acquisition.

Some examples of these systems are - [4] have explored teaching formulaic expressions using a scenario based interactive environment where the users are presented with videos and are able to enter their answers to questions in the videos. Whereas, [5] have explored CAVE environments to teach language through immmersion. Authors like [6] have explored other immersion techniques such as Second Life and developed a task-based learning paradigm using it. While [7] have reviewed four immersive technologies, some of which combine the use of AI and immersion and understand speech-based input in order to deliver a richer and naturalistic conversation experience

We see that most of the applications built are philosophically following the "instructors use what they have" paradigm like [8] explores the use of available internet tools and how it aids language teaching. However, it is hard to find a perspective that first drove the requirements from a language teacher on what would help their class - this is one of the novel points of this work.

In addition to virtual immersive worlds being important, it is also duly noted that the embodied agent domain has significantly contributed to the language learning experiences of students. [9] have reviewed such agents and its effects on speech correction for students and we take inspiration from them for enhancing our embodied virtual agent (described later) in the future.

DOI: https://doi.org/10.3217/978-3-85125-609-3-13

From the literature review, we see that although multi-modal learning is explored in how information is presented to the students. It hasn't being explored too much as a means of input to the system. That is, systems do not exist out there in the language teaching domain that can interpret multiple simultaneous modes of input (such as gesture and speech) as also noted by [10] who points out that gestures are important in communication and language learning. [10] herself has built a system that takes in gesture input from the learners in order to move the characters in the virtual world such as picking up an object, etc. However, there is another component of multi-modal input that has yet to be explored. We know that natural deictic expressions wherein speech and gestures are used together to communicate are common. This is especially seen while communicating in a new language that is being acquired because, the speakers sometimes fall short of vocabulary necessary to communicate. We haven't been able to find a system that can take into account such a multi-modal dialogue as input to teach foreign language and is another novel point of this work.

We also see that most of the applications that are in the space of simulating reality do not explore human-scale immersion or non-intrusive device based user-experiences augmented with AI which we believe provides a more naturalistic experience.

Therefore, we justify that a system is needed that does the following

- Allow a conversational (speech based) experience
- Understands naturalistic conversation that involves deictic speech
- Requires no hardware to be worn by the user
- Is rich in visual context and culture by using human-scale immersion

The above points guide us to simulate a real-life experience of traveling to a foreign country - a type of immersion highly hailed in the community. However, it needs to additionally be designed to meet some pedagogical goals outlined in section 4. In the next section, we briefly describe the system itself and how it aims to satisfy the above points.

3 Achieving Natural Realism

This section sets up our readers with the context for the next two sections by giving a brief idea of the user experience.

We have used the Cognitive Immersive Room (CIR) which is set-up as a 360 degree panoramic screen that surrounds a group of users as seen in fig. 1 and its computer generated representation in fig. 2. The display has a diameter of 12m and height of 3.8m.

For the context of this work, the interaction can be imagined as - the users can come in into a simulated Chinese restaurant and talk to the AI Agent (embodied as a Panda on the screen) in the context of a typical restaurant experience using lapel-microphones and gestures (in case of deictic remarks). The panda is able to understand what the users are saying and is able to respond (through actions, facial reactions, images and speech) appropriately. We urge more enthusiastic

DOI: https://doi.org/10.3217/978-3-85125-609-3-13

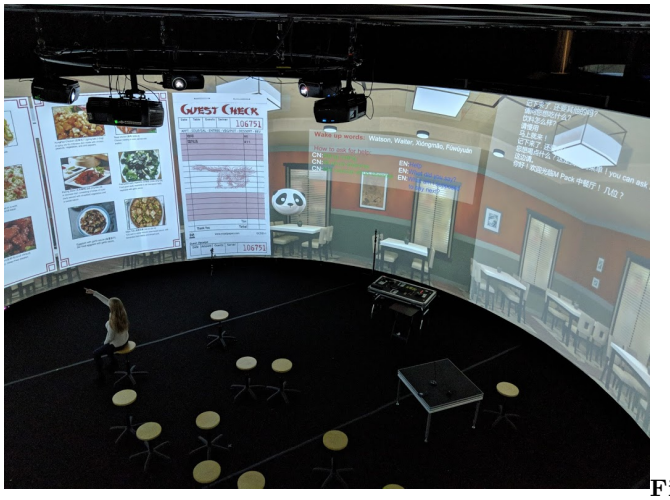

Fig. 1. Interaction in the immersive restaurant (top-view)

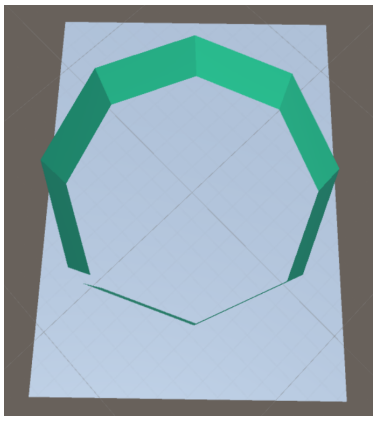

Fig. 2. Computer render of the CIR display

readers to see [11] who have explained in detail the exact interaction that is possible. Additionally, a short demo of the system can be seen here[3]. Such a system can be used in foreign language learning classes in addition to classroom instruction so that students get a chance to role-play in a realistic scenario with context (e.g. visual, audio, etc.)

4 Pedagogical Requirements

From the literature review, we can see that the use of VR/AR, intelligent learning tools and environments mainly focuses on the "tool" or "tutoring" features of the emerging technologies in foreign language teaching and learning. It lacks of interactivity or communication in context, that is, the sense of immersion in real-life situation. In addition, all designs are technology-driven. Seeing the shortcomings of existing design and/or research, the current work started with the collaboration of designers, programmers and language teaching experts since day one of the research project. The technical designers (e.g., designers and programmers) asked a fundamental question to the instructor of the Mandarin Chinese course: What pedagogical goals or requirements you would like to see in the immersive classroom? From the instructor (i.e. language acquisition expert), an ideal cognitive immersive classroom would provide a near real-life experience to the students so that they could interact with an AI agent for task-based learning. Specifically for the pilot scene, we chose ordering food as a task to be completed in Mandarin Chinese in an immersive restaurant. In this restaurant, the students are expected to interact with the AI Waiter, from greeting to ordering

[3] https://www.youtube.com/watch?v=lZWtDqhFlAc

DOI: https://doi.org/10.3217/978-3-85125-609-3-13

food, to consulting and clarifying food items and requesting a check. Please note that the system is designed for students who have learned the vocabulary and sentence structures required to complete these tasks through their classroom instruction. The pedagogical requirements for the task-based, communicative language learning are:

1. Collaborative Learning: Students work as a pair or groups to help with each other in the learning process
2. Task-Based Language Teaching (TBLT): Students need to complete the food-ordering task which differs from most other existing research into VR/AR immersive Mandarin language learning
3. Communicative: Students have opportunities to communicate with the AI agent like how they would communicate in a real-world
4. "Interpersonal" communication: Students interact with the AI waiter as if they were customers at a Chinese restaurant
5. Noticing and Awareness Raising: Students need to be alerted for errors and correct their errors for the success of a communicative task.
6. Scope of Language: The technology considers the limited knowledge of the students and help raise it through the dialogue
7. Culture: The system should consider that culture presentation helps aid language acquisition.
8. Help: Students need to have tools (e.g., visual, audio, and/or textual) to help or assist them when they are stuck in the process of task completion.

5 Technology that Supports Pedagogical Goals

Given the goals from the language teaching expert, we set-out to build a system with the following technology specifications and show how they map to each goal. We use *#number* notation to refer to a point from the previous section.

1. Enable Visually-Rich environment: We employed the 360 degree panoramic screen seen in fig. 1 to immerse the students in a restaurant scene with a rich 3-D model of a Chinese restaurant. This aids part of requirement #3 and #4 by simulating the "real world" and provides the visual context for where this communication needs to happen.
2. Minimal use of wear-able devices - In order to encourage collaborative learning (#1), we want the students to be able to look at each other and communicate with each other freely in the "restaurant". This becomes difficult with technologies that employ headsets or other intrusive devices. Thus, we ask the students only to wear a lapel microphone in order to communicate with the system. This has motivated the development efforts of human-scale immersion and computer vision to recognize gestures. It adds to the realism of the environment by allowing a free, natural and interpersonal interaction (#3, #4), and allows for collaborative learning.
3. Voice-Based Dialogue: We use speech-to-text(STT) and text-to-speech (TTS) to have our agent *listen* and *talk* to the students in a natural fashion(#4).

DOI: https://doi.org/10.3217/978-3-85125-609-3-13

We additionally use the Watson Conversation Assistant to have the agent *understand* what is being said and pick appropriate responses. The STT expects near-perfect pronunciations otherwise it would misunderstand what was said and lead to the agent being in a confused state. This enables TBLT using speech as the agent responds to each task successfully completed using speech(#2) (e.g. greeting, ordering food, etc.) and also allows for interpersonal communication (#3).

4. Transcription: In order to give feedback on the pronunciation of the students, they were at all times shown the transcription of what they said in real time. This was displayed on the screen and idea was that the students will be able to identify exactly which word was misunderstood and repeat their sentence with focus on that word raising their awareness (#5) of their pronunciation on specific pinch-points.
5. Dialogue Creation - In order to keep the communication natural, we do not give the students a script to follow. Instead, we use the agent-initiated dialogue paradigm where the agent always initiates the next turn and asks specific questions that can be answered by the students. This way, we get to achieve three objectives: stick to tasks that can be accomplished using the language level of the students (#2) (e.g. students cannot accomplish tasks like getting a take-out at the current language level and the agent could avoid that dialogue branch); elicit certain responses from the students chosen by the language teacher as a general tough spot and; use some out-of-scope vocabulary to raise the level controllably (#6). Such a dialogue is designed to be interpersonal e.g. students could ask "what kind of tea do you have?" and talk directly to the waiter (#3,#4). Each dialogue turn could be considered as a task that the student needs to complete (e.g. ordering food) and is communicative through speech. Each dialogue turn can be spoken in Mandarin or English (L1) to help the students if they are stuck (#8).
6. Multiple output responses - The agent has a list of sentences for each meaning it wants to convey to the student and it randomly chooses from the list to respond to the student (#2). For example, the agent could randomly choose between asking "What would you like to drink?" or "Do you want a drink?" - both, achieving the same outcome. All alternative sentences are designed keeping in mind the students' expected understanding of the language. This is done to make the system less repetitive and give the students more exposure to language in context to the task (#6).
7. Gesture (pointing) recognition: One of the observations of natural dialogue and completing tasks is using deictic speech, especially while learning foreign languages. Using our sensor fusion research, we were able to combine multi-modal input in order to understand deictic sentences like "I would like to order *this*" (#2) where *this* is resolved by where the student was pointing on the screen (#3,#4). This multi-modal communication allows for affordances when a student does not know how to pronounce something on the menu (#6,#8). As the goal of the interaction is to have minimal wearable devices, we have used computer vision technologies to recognize gestures.

DOI: https://doi.org/10.3217/978-3-85125-609-3-13

We encourage readers to see [12] and [13] to read more about how this is enabled.

8. Listening and learning: The AI agent speaks at a standard pace (we did not slow it down as it is expected of the students to learn to understand the standard pace #2). The panda might also use some out-of-vocabulary words (which were carefully placed in certain dialogue turns #6). We expect the students to learn from this experience and, to do so, they could ask the agent questions like "What did you say?" or "Can you repeat that?" (#8) The agent uses pre-coded responses to explain what it said.
9. Story-telling - Chinese cuisine has a rich history and culture behind it (#7). On being asked, the agent can provide an interactive computer-generated narration for the items on the menu. This interactive narration (#3,#4) is presented in L1 given the limited knowledge of students' L2. To achieve this, we have integrated technology from [14].
10. Assessment Module Post-role play, the panda can give feedback to the students in the following ways (#5,#8)

    – Show a list of tasks students did not complete in Mandarin (i.e. tasks skipped by using L1 or deictic speech) - both an indicator of lack of language execution. This list provides the teacher and the students with opportunities to improve the students' language (#2). The students can practice sentences related to the tasks skipped with the AI Agent.

    – Pitch-contour: Since Mandarin is a tonal language, at the end of the module, students can practice the tone of individual words with the Agent. They can do this by comparing the graph of their own pitch-contour against the graph of the ideal pronunciation in real time and, hearing the correct pronunciation from the Agent.

Table 1 summarizes the mapping of the technology design to the pedagogical goals. Pedagogical goal 8 i.e. providing with the right amount of help to the students so that the role-play can go on in this new interaction paradigm is a greater challenge and has been further discussed in detail in [11].

Table 1. Summary of pedagogical goals mapped to technology design

Pedagogical Goal	**Technology Design**
Collaborative Learning	2
TBLT	3,5,6,7,8,10
Communicative	1,2,3,5,7,9
Interpersonal Communication	1,2,3,5,7,9
Noticing and Awareness Raising	4,10
Scope of Language	5,6,7,8
Culture	9
Help	5,7,8,10

DOI: https://doi.org/10.3217/978-3-85125-609-3-13

6 Technology Implementation

This section gives a brief overview of the implementation of the technology design discussed in the previous section.

Visuals: We use Unity Game Engine[4] to drive the rich visual output on the massive panoramic screen. Special software can warp the displays and achieve the effect of panoramic vision.

Gestures: Commonly used wearable devices that recognize gestures are replaced with computer vision systems that can "see" the user and recognize gestures. We mainly use the skeletal tracking from Kinects and software on top of it to identify gestures from the skeletons. This has been throughly described by [12].

Speech: We use IBM Watson Speech-to-text[5] and Conversation service [6] to classify the verbal input into one of the pre-defined intents designed by us. The intent recognized in this stage is used to drive the interaction. E.g. if intent recognized was *greeting*, the interaction would be that the panda moves to the host-stand and says "hello" to the students. We use Baidu services for text-to-speech generation[7].

Many of the above discussed technologies are a research area of its own. Additionally, a massive engineering effort has gone into integrating them especially with sensor-fusion. It is hard to describe all the details in this short work which focuses on outlining the technology design principles that were followed to justify pedagogical requirements.

However, [13] have outlined the integration engineering efforts giving a detailed overview of the implementation of all technologies involved, while [12] have specifically outlined the workings of the computer vision systems. A combination of these two works will give enthusiastic readers a good insight into the implementation of the technology.

7 Pilot User Study

The aim of this work is to build technology to support the pedagogy. While doing so, the first step in building a new interaction paradigm is to measure its usability and likability. In order to do that, we conducted a pilot study discussed here briefly to show that the interaction was usable and likable. Through a series of testing and feedback over several months, we were able to improve the interaction by identifying interaction challenges and designing solutions to them that are described in detail by [11].

In the eventual formal user study, we invited 16 students from the class of the collaborating professor who teaches Chinese-1 at the same university. We asked them to interact with the system in Mandarin in the food-ordering context. We

[4] https://unity3d.com/
[5] https://www.ibm.com/watson/services/speech-to-text/
[6] https://www.ibm.com/watson/services/conversation/
[7] yuyin.baidu.com/tts

DOI: https://doi.org/10.3217/978-3-85125-609-3-13

observed that the students were able to complete most of the tasks showing that the system is usable. We asked them several questions post-interaction to evaluate how natural the system was as this hints towards how well we were able to immerse the students. All of the students indicated that the liked that they were able to gesture and point to menu items (i.e. deictic speech). On the subject of "Did the immersive restaurant feel realistic? Why or why not?", the overall sentiment was positive with at least 10 people strongly indicating that they liked the system. The general complaints were with respect to the floating head of the panda and the lack of leniency in speech-recognition. This encourages us to further enhance the system. Our biggest encouragements were feedback on the lines of the students being able to learn a new word simply through the interaction and the context that was provided to them. The exact pedagogical benefit of this study is still work-in-progress but the positive comments hint towards success. Additional analysis of the user-study can be found in [11].

8 Conclusion and Future Work

We have presented a prototype version of a system that was built to support specific goals of TBLT in Mandarin Chinese language learning for college students. This system is a multi-disciplinary collaboration between educators, language learners, computer scientists, electrical engineering etc. Our pilot studies show a promising future for this project and we plan to scale this to multiple scenes having multiple agents to support different roles and types of dialogues.

9 Acknowledgment

This work is supported by the Cognitive and Immersive Systems Laboratory (CISL), a collaboration between IBM Research and Rensselaer Polytechnic Institute, and also a center in IBM's AI Horizon Network. The authors would like to thank all of the researchers at CISL and the colleagues at IBM Yorktown Heights and IBM China for their contributions to this research. Without their help, this project would not have been possible.

References

1. Lin, T.-J. & Lan, Y.-J. Language Learning in Virtual Reality Environments: Past, Present, and Future. *Educational Technology & Society* **18,** 486–497 (2015).
2. Blyth, C. Immersive technologies and language learning. *Foreign Language Annals* (2018).
3. Sankey, M., Birch, D. & Gardiner, M. *Engaging students through multimodal learning environments: The journey continues* in *Proceedings ASCILITE 2010: 27th Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education: Curriculum, Technology and Transformation for an Unknown Future* (2010), 852–863.

DOI: https://doi.org/10.3217/978-3-85125-609-3-13

4. Taguchi, N., Li, Q. & Tang, X. Learning Chinese Formulaic Expressions in a Scenario-Based Interactive Environment. *Foreign Language Annals* **50,** 641–660 (2017).
5. O'Brien, M. G., Levy, R. & Orich, A. Virtual immersion: The role of CAVE and PC technology. *Calico Journal* **26,** 337–362 (2009).
6. Lin, T.-J., Wang, S.-Y., Grant, S., Chien, C.-L. & Lan, Y.-J. Task-based teaching approaches of Chinese as a foreign language in Second Life through teachersfffdfffdfffd perspectives. *Procedia Technology* **13,** 16–22 (2014).
7. Lane, H. C. & Ogan, A. E. *Virtual environments for cultural learning* in *Second Workshop on Culturally-Aware Tutoring Systems in AIED 2009 Workshops Proceedings* (2009).
8. Yang, S. C. & Chen, Y.-J. Technology-enhanced language learning: A case study. *Computers in Human Behavior* **23,** 860–879 (2007).
9. Wik, P. & Hjalmarsson, A. Embodied conversational agents in computer assisted language learning. *Speech communication* **51,** 1024–1037 (2009).
10. Si, M. A virtual space for children to meet and practice Chinese. *International Journal of Artificial Intelligence in Education* **25,** 271–290 (2015).
11. Divekar, R. R. *et al. Interaction Challenges in AI Equipped Environments Built to Teach Foreign Languages Through Dialogue and Task-Completion* in *To appear in Proceedings of the Designing Interactive Systems Conference* (2018). To appear.
12. Zhao, R. *et al. An Immersive System with Multi-modal Human-computer Interaction* in *The IEEE Conference on Automatic Face and Gesture Recognition (FG)* (May 2018).
13. Divekar, R. R. *et al. CIRA—An Architecture for Building Configurable Immersive Smart-rooms* in *To appear in Proceedings of Intellisys 2018* (2018). To appear.
14. Si, M., Battad, Z. & Carlson, C. *Intertwined Storylines with Anchor Points* in *Interactive Storytelling: 9th International Conference on Interactive Digital Storytelling, ICIDS 2016, Los Angeles, CA, USA, November 15–18, 2016, Proceedings 9* (2016), 247–257.