ICA 2022
24th INTERNATIONAL CONGRESS ON ACOUSTICS
October 24(Mon) - 28(Fri), 2022 Gyeongju, Korea

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

A spatially-aware companion system for language learning and foreign-language dialogue

Albert Chang[(1)], Samuel Chabot[(2)], Jonathan Mathews[(3)], Shannon Briggs[(4)], Tomek Strzalkowski[(5)], Mei Si[(6)], Jonas Braasch[(7)]

[(1)]School of Architecture, Rensselaer Polytechnic Institute, USA, changa4@rpi.edu
[(2)]School of Architecture, Rensselaer Polytechnic Institute, USA, chabos2@rpi.edu
[(3)]School of Architecture, Rensselaer Polytechnic Institute, USA, mathej4@rpi.edu
[(4)]Cognitive and Immersive Systems LaboratoryRensselaer Polytechnic Institute, USA, briggs4@rpi.edu
[(5)]Cognitive Science, Rensselaer Polytechnic Institute, USA, tomek@rpi.edu
[(6)]Cognitive Science, Rensselaer Polytechnic Institute, USA, sim@rpi.edu
[(7)]School of Architecture, Rensselaer Polytechnic Institute, USA, braasj@rpi.edu

ABSTRACT
Dialogue systems have become a popular research medium as recent advances in task-oriented and open-domain systems combined with deep learning technologies have increased the potential for practical applications across many disciplines. One such vein of applications involves multi-modal dialogue systems deployed in interactive spaces that seek to provide an immersive experience for participants. This project proposes a combination of spatial awareness with a multi-modal, immersive dialogue system as a potential interactive medium to provide an additional layer of immersion. The system employs an array of audio/visual sensors that track participants within the interactive space. It responds contextually depending on the application and information domain, for example, by displaying and sonifying conversational agents at accurate spatial locations. The current application of this system involves Mandarin language learning, in which the system will act as both a learning medium and conversation augmentation system to provide students with an immersive environment to learn a language and provide real-time feedback during the learning process. This project aims to provide insight into interactive spaces for education and general conversation applications and demonstrate the capabilities of combining spatial awareness with a multi-modal dialogue system.

Keywords: Dialogue, Immersion

1 INTRODUCTION

Virtual reality systems found its place in education and training scenarios, with studies finding significant benefits over traditional learning methods [26, 20, 14, 19, 15, 23]. A remaining problem for long-term training is the effect of cybersickness when using head-mounted displays [21, 34]. A recent study showed that even when watching a 3D movie, about 55% of the viewers complain about the side effects [30]. As an alternative approach, Rensselaer developed, erected, and operates two large-scale collaborative, immersive cognitive laboratories with panorama screens, the CRAIVE-Lab [5, 3, 28] and the EMPAC Panorama Screen system [4, 12, 11]. We define *collaborative, immersive cognitive systems* as environments where:

1. groups of humans can communicate naturally with each other without obstruction from wearable devices *(collaborative)*,
2. groups of people are embedded in a panoramic display with surround sound capabilities *(immersive)*,
3. users can draw from intelligent computing capabilities *(cognitive systems)*.

Both labs can serve as classrooms providing access up to 49 students while avoiding cybersickness. The CRAIVE-Lab has a useable floor space of 12 $\times$ 10 sqm (Fig. 1); the Panorama Screen has a diameter of 12 m. The labs provide a new form of embodied learning, one where the learned material no longer needs to be scaled to the real world because situated learning is very similar to the real world experience. One no longer needs to recall what was written in the textbook for what to say to the Chinese customs officer because one learned the appropriated phrases in an immersive, human-scale environment talking to a life-size interactive avatar presented on screen [11].

The theories of embodied cognition suggest that people naturally construct and utilize environmental cues to help them reason and lower cognitive workload [1, 9, 16]. An important feature of working in an immersive system is the affordance of using body movements and gestures that are intuitive to communicate and perform tasks and therefore leave the interface transparent. For example, the user can "grab" and "drag" new information and "throw away" irrelevant items. In contrast, most existing computer-based learning systems require the users to sit in front of a computer and use a keyboard and mouse to interact. The users' body movements are generic for operating a computer and have nothing to do with their tasks. Immersive cognitive systems thus can allow users to interact with the digital environment in a similar way as they interact with the physical world. By turning the meaningless hand movements of mouse click and pulling down a drop-down menu into something consistent with the operation's effects, we expect immersive cognitive systems to be more engaging and natural to use.

This paper describes a method of using a dialogue system with an immersive educational environment to facilitate interactions between students and automated services, for example, a tutor avatar that teaches a student vocabulary. The immersive dialogue system for educational use should possess the following components: (i) A dialogue system that receives input from the user, decides on a response, and sends the response to the appropriate visual and audio output systems, (ii) Audio and visual hardware to receive input from participants and output their respective responses, (iii) A database/third-party resource to draw answers from regarding the domain(s) of interest, (iv) A spatial tracking system to identify where participants are in the space and identify individual participants, and (v) A visual avatar system that is displayed to the participants.

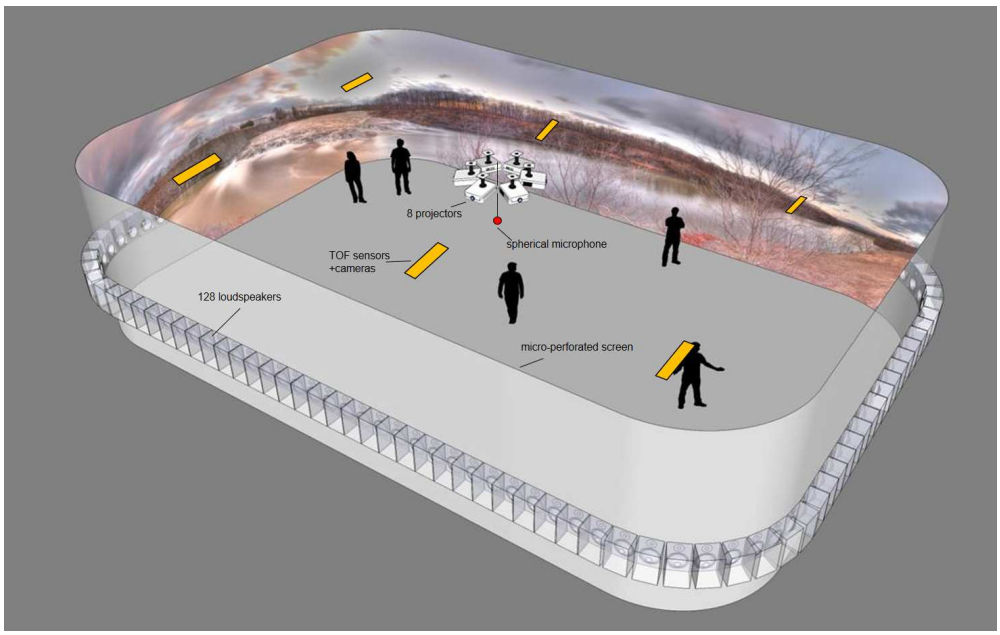

Figure 1. CAD model of the CRAIVE-Lab (usable floor area: 12$\times$10 sqm).

2 IMMERSIVE SYSTEM INFRASTRUCTURE

This project was developed in the Collaborative-Research Augmented Immersive Virtual Environment Laboratory (CRAIVE-Lab) at Rensselaer Polytechnic Institute – [6]. The lab addresses the need for a specialized virtual-reality (VR) system for the study and enabling of communication-driven tasks with groups of users immersed in a high-fidelity multi-modal environment located in the same physical space (Fig. 2). It is therefore ideal for the integration of a dialog system. For the visual domain, an eight-projector front-projection display to (re)create scenes on a seamless screen has been created. A DMX-controlled lighting system can tune the color and intensity of the interior light, which illuminates the participants without much spill to the screen area. For the acoustic domain, a 134-loudspeaker-channel system has been designed and installed for Wave Field Synthesis (WFS) with the support of Higher-Order-Ambisonic (HoA) sound projection to render inhomogeneous acoustic fields. The screen material is micro-perforated to minimize acoustical reflections. An intelligent position-tracking system estimates current user locations and head orientations as well as positioning data for other objects. For the tracking system, a hybrid visual/acoustic sensor system is being used to emulate the humans' ability to extract robust information by relying simultaneously on different modalities. A network of six cameras has been installed in CRAIVE-Lab as well as a time-of-flight sensor array using six Microsoft Kinects. A 16-channel spherical ambisonic microphone with additional peripheral microphones is used for acoustical tracking – see Fig. 2.

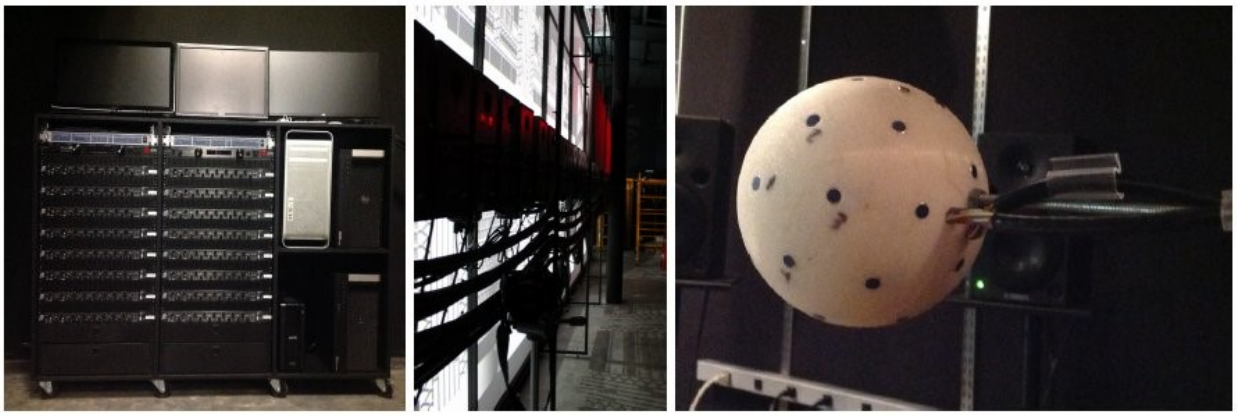

Figure 2. Audio system of the CRAIVE-Lab, from left to right: Audio rack, loudspeaker system, spherical microphones.

The CRAIVE-Lab hosts several projects including the Mandarin Project [10], which aims to teach students Chinese language in an immersive way. Several scenes have been developed for the project including the restaurant scene shown in Fig. 3.

Figure 3. Scene from the Mandarin Project Class. The restaurant environment was created using Unity.

3 SPATIAL TRACKING

In order to localize participants, an acoustic tracking system has been developed and implemented into the CRAIVE-Lab [25, 24]. The audio tracking system uses a 16-channel spherical microphone array in conjunction with a sparse iterative beamforming algorithm to enable low-latency estimation of acoustical sound sources at a low computational cost. Our Sparse Iterative Search (SIS) method starts out by analyzing the whole spherical area with equal-area grid size. The algorithm computes the received energy in each area and contracts the analyzed area and grid size based on the received energy. Using an iterative approach, the algorithm zones into the energy-emitting sound sources with high accuracy due to the small grid size. Incorporating conditions based on temporal smoothing and diffuse energy estimation further refines this process. This way, the algorithm can track up to four simultaneous static or moving sound sources.

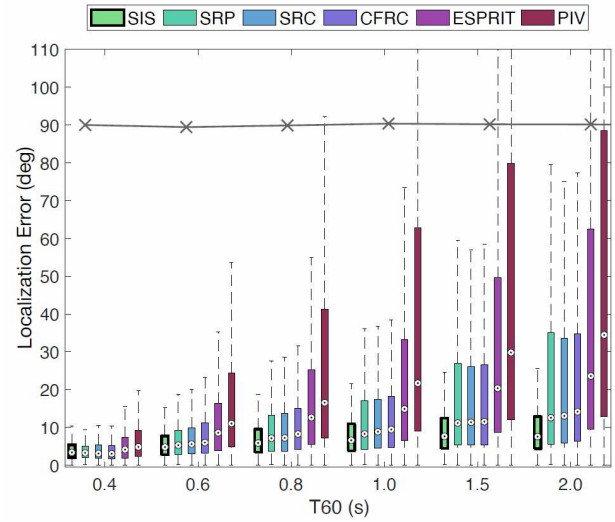

Figure 4. Localization errors for different audio tracking algorithms as a function of reverberation time – from [24].

The system was successfully tested against existing algorithms, including Coarse-to-Fine Region Contraction (CFRC), Eigenbeam Estimation of Signal Parameters with Rotationally Invariant Techniques (EB-ESPRIT),

Pseudo-intensity Vector (PIV), Sparse Iterative Search (SIS), Stochastic Region Contraction (SRC), and Steered Response (SRP). The Sparse Iterative Search (SIS) method maintains an average localization error of about 8° at 2 seconds of reverberation time (T_{60}), while all other methods showed errors of more than 12°, typically at a much higher computational cost – see Fig. 5 and [24].

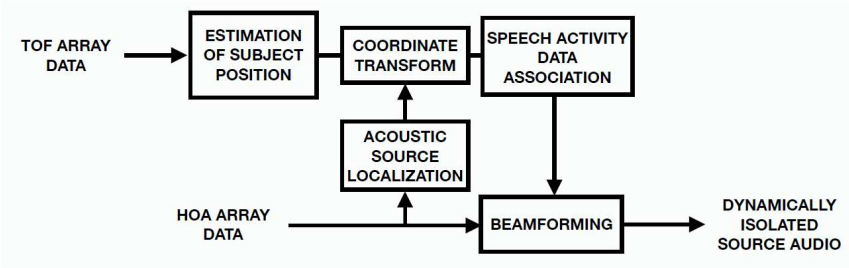

Figure 5. Architecture of the integrated audio-visual tracking/speech isolation system – from [24].

The audio tracking system has been integrated with a time-of-flight sensor-based tracking system as shown in Fig. 5. The hybrid tracking system received input from the spherical microphone (label: *"HOA ARRAY DATA"*) using the higher-order ambisonics (HoA) format and six time-of-flight (ToA) sensors (Microsoft Kinect), see label: *"TOF ARRAY DATA"*. The acoustic data is then processed through the previously described Sparse Iterative Search (SIS) method – see box *"ACOUSTIC SOURCE LOCALIZATION."* The time-of-flight sensor data is analyzed using a Gaussian mixture model with background subtraction [33, 2, 27, 29, 18] – see box *"ESTIMATION OF SUBJECT POSITION."* Given that the ToF data is processed using a Cartesian Coordinate system and the HoA data using a polar coordinate system, a transformation has to be applied to bring both data sets into the same coordinate system – see box *"COORDINATE TRANSFORM."* Here, it is also noteworthy that the angles of the acoustic source signal can be much better determined than the distance of the source from the spherical microphone. The additional ToF data partially resolve the distance uncertainty. Alternatively, one could also use a second HoA array to triangulate the sound source. Detected utterances are assigned to the estimated location of the sound source (box: *"SPEECH ACTIVITY DATA ASSOCIATION"*) and both can be used used to steer a beamformer (box: *"BEAMFORMING"*). For details see [25].

Using this multi-modal approach, the participants can be tracked optimally. When a participant is talking, we can identify the source of the utterance and assign this utterance to the visually tracked object. We can continue to track the participants when moving in quiet. By tracking participants acoustically, the tracking process is also less susceptible to difficult lighting conditions, although the latter can be adjusted automatically using a DMX lighting system. Under certain conditions, the participants benefit from a dark interior, for example, when being presented with a movie or panoramic photography/videography.

4 DIALOGUE SYSTEM

4.1 Background

Dialogue systems involve human-computer conversations with at least one human and one computer system. The applications of dialogue systems are ubiquitous, such as personal assistants available on smartphones such as Google Assistant and Siri, non-player characters in video games, customer service bots that handle frequently asked questions, and chatbots that employ machine learning to hold casual conversations. All of these systems fundamentally use either or a combination of pre-defined choices and Natural Language Understanding as inputs and employ a dialogue logic system to determine the best course of action in responding. Using their application as a basis, dialogue systems can be categorized into two groups: task-oriented and non-task-oriented systems [8].

Task-oriented systems, akin to their name, aim to assist users with a task of some sort, whether that be booking travel tickets or finding products. These systems focus on a goal or set of goals to accomplish by the

end of the conversation and may have additional mitigation to minimize off-topic conversation and task failure. These systems typically consist of Natural Language Understanding (NLU), dialogue state tracking, policy learning, and Natural Language Generation (NLG). In a nutshell, each component – respectively – is responsible for parsing inputs, predicting the user's goal, generating the best system action, and finally responding with natural language generation – e.g., see [8].

Non-tasked-oriented systems, on the other hand, focus on conversing on open domains where the conversation is steered by the human(s) interacting with the system. To accomplish open-domain discussion, non-tasked-oriented systems use a retrieval-based, generative, or hybrid approach to dialogue responses [31]. Retrieval-based systems can employ more intelligent and fluent responses [17], while generative systems allow for more context-aware responses that are not part of the original corpus [32].

4.2 Dialogue System Architecture

The immersive dialogue system consists of a three-step process that handles visual and audio inputs and outputs separately but uses both to provide the most amount of context needed to best respond in the current dialogue. First, the systems uses a speech-to-text system, gesture/body language interpreter, and tracking system to handle inputs, processes the inputs in the dialogue engine. Second, working with the game engine in tandem, the dialogue engine gathers information from third-party sources and an internal database and decides on a combined audio/visual response. Finally, the response is carried out through the avatar system and text-to-speech output. The design of the immersive dialogue system builds upon the work done by Divekar et al. [13] for the initial prototypes for the Mandarin Project by using the core architecture and extends its functionality to include auditory spatial tracking enhancements in addition to a virtual human avatar that can act as both the subject of a conversation and an assistant to group conversations.

Figure 6 shows the dialogue system design for the immersive dialogue agent that also includes an avatar for visual representation. See also Fig. 7, which shows students interacting within a garden scenario in the Mandarin Project where they can talk to agents to participate in various activities, such as one that will teach basic Tai chi (white-clothed avatar pictured to the right).

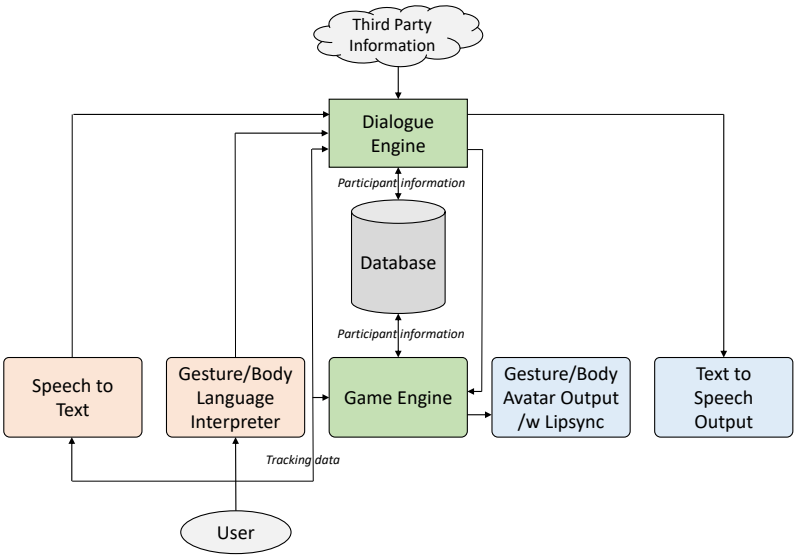

Figure 6. Dialogue System Architecture.

Visual and audio components are utilized as described in the Immersive System Infrastructure and Spatial Tracking sections to track participants as they use the system. The additional functionality of agent gaze can be implemented as a standalone application using cameras that also takes in inputs from the visual and audio

tracking system, displays an image of an agent, and adjusts the agent's gaze to look towards the head of the active participant. Identifying the active speaker requires both a visual identification in the installation space along with the participant currently or most recently speaking.

A primarily task-oriented system was decided upon as the basis for the dialogue system due to its intended purpose of acting as an assistant to language learning and foreign-language dialogue. The unpredictable and demanding nature of multi-user interactions with real-world conversations also proposes a unique challenge for a multi-modal system, which makes the inherently unpredictable nature of generative dialogue systems less suitable for this application. Furthermore, there are currently no standardized designs nor testing procedures for comparison – as such, determining the effectiveness renders generative-leaning models too unpredictable for evaluating initial user feedback to a system intended to educate or assist.

Figure 7. Students interacting with the Mandarin dialogue system.

5 DISCUSSION & OUTLOOK

The current application of the immersive system with the Mandarin Project demonstrates the potential reach and impact through enabling potential content creation by such a system. A successful pilot program has been conducted using the space to supplement an undergraduate Chinese class held at Rensselaer Polytechnic Institute [10]. Both student feedback and learning performance were notably positive, which shows promising results for immersive educational content. Future work for the project aims to integrate the Mandarin Project as a formal addition to the Chinese curriculum at Rensselaer with expanded content covering more topics and interactable scenarios.

Additional future work will focus on a practical demo involving a game of *"I Spy"* where participants will locate objects or points of interest in an environment by responding in Mandarin and/or pointing at an item. The demo will employ additional spatial tracking features to enhance immersion while playing the game by building upon the panoramic imagery system from Chatbot et al. [7]. In addition, an investigation will be done on incorporating the additional features of the new spatial tracking system into the dialogue, as adding multi-modal components has been demonstrated to improve generative-based dialogue systems. In this context, Liao et al. combined image data with text in the fashion domain and found their hybrid model outperformed other state-of-the-art models [22].

The potential for other applications outside of the education sector cannot be understated either, as potential use cases such as remote monitoring and collaboration for business applications or site experiences for tourism can exploit the spatial awareness and group engagement the immersive system allows. In contrast to virtual reality, the immersive system can also allow for a safer alternative for those experiencing motion sickness as there is less disconnect between the person and visual feedback from the system.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant No. HCC-1909229, the United States Air Force under Contract No. FA8750-21-C-0075 and the IBM Corporation under the Artificial Intelligence Research Collaboration Agreement No. W1771793 between IBM and Rensselaer. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NSF, USAF, or IBM Corporation. The research presented in this paper was carried out within Rensselaer Polytechnic Institute's Cognitive and Immersive Systems Laboratory.

REFERENCES

[1] L. W. Barsalou. Grounded cognition. *Annu. Rev. Psychol.*, 59:617–645, 2008.

[2] I. Bhattacharya, M. Foley, N. Zhang, T. Zhang, C. Ku, C. Mine, H. Ji, C. Riedl, B. F. Welles, and R. J. Radke. A multimodal-sensor-enabled room for unobtrusive group meeting analysis. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 347–355, 2018.

[3] J. Braasch, J. Carter, S. Chabot, and J. Mathews. An immersive teleconferencing system using spherical microphones and wave field synthesis. In *22nd International Congress on Acoustics*, pages Paper ICA2016–826, Buenos Aires, Argentina, September 2016.

[4] J. Braasch, J. Goebel, and T. Vos. A cinematic spatial sound display for panorama video applications. *Organised Sound*, 15(3):260–270, 2010.

[5] J. Braasch (PI), R. Radke (Co-PI), B. Cutler (Co-PI), J. Goebel (Co-PI), and B. Chang (Co-PI). Mri: Development of the collaborative-research augmented immersive virtual environment laboratory (craive-lab), 2012–2015. NSF #1229391.

[6] S. Chabot and J. Braasch. Interactive application to control and rapid-prototype in a collaborative immersive environment. In *Audio Engineering Society Convention 151*. Audio Engineering Society, 2021.

[7] S. Chabot, J. Drozdal, M. Peveler, Y. Zhou, H. Su, and J. Braasch. A collaborative, immersive language learning environment using augmented panoramic imagery. *Proceedings of 6th International Conference of the Immersive Learning Research Network, iLRN 2020*, pages 225–229, 6 2020.

[8] H. Chen, X. Liu, D. Yin, and J. Tang. A survey on dialogue systems: Recent advances and new frontiers. 11 2017.

[9] A. Clark. *Supersizing the mind: Embodiment, action, and cognitive extension*. OUP USA, 2008.

[10] R. R. Divekar*, J. Drozdal*, S. Chabot*, Y. Zhou, H. Su, Y. Chen, H. Zhu, J. A. Hendler, and J. Braasch. Foreign language acquisition via artificial intelligence and extended reality: design and evaluation. *Computer Assisted Language Learning*, pages 1–29, 2 2021.

[11] R. R. Divekar, J. Drozdal, Y. Zhou, Z. Song, D. Allen, R. Rouhani, R. Zhao, S. Zheng, L. Balagyozyan, and H. Su. Interaction challenges in ai equipped environments built to teach foreign languages through dialogue and task-completion. In *Proceedings of the 2018 Designing Interactive Systems Conference*, DIS '18, pages 597–609, New York, NY, USA, 2018. ACM.

[12] R. R. Divekar, M. Peveler, R. Rouhani, R. Zhao, J. O. Kephart, D. Allen, K. Wang, Q. Ji, and H. Su. CIRA: An architecture for building configurable immersive smart-rooms. In *Proceedings of SAI Intelligent Systems Conference*, pages 76–95. Springer, 2018.

[13] R. R. Divekar, M. Peveler, R. Rouhani, R. Zhao, J. O. Kephart, D. Allen, K. Wang, Q. Ji, and H. Su. Cira: An architecture for building configurable immersive smart-rooms. pages 76–95, 2019.

[14] H. A. El-Mounayri, C. Rogers, E. Fernandez, and J. C. Satterwhite. Assessment of stem e-learning in an immersive virtual reality (VR) environment. American Society for Engineering Education, 2016.

[15] E. Fokides and P. Atsikpasi. Factors affecting primary school students' learning experiences when using muves: Development and validation of a scale. In *Integrating Multi-User Virtual Environments in Modern Classrooms*, pages 185–206. IGI Global, 2018.

[16] S. Goldin-Meadow and M. W. Alibali. Gesture's role in speaking, learning, and creating language. *Annual review of psychology*, 64:257–283, 2013.

[17] Z. Ji, Z. Lu, and H. Li. An information retrieval approach to short text conversation. 8 2014.

[18] L. Jia and R. J. Radke. Using time-of-flight measurements for privacy-preserving tracking in a smart room. *IEEE Transactions on Industrial Informatics*, 10(1):689–696, 2013.

[19] S. H. Kidd and H. Crompton. Augmented learning with augmented reality. In *Mobile Learning Design*, pages 97–108. Springer, 2016.

[20] P. T. Kovács, N. Murray, G. Rozinaj, Y. Sulema, and R. Rybárová. Application of immersive technologies for education: State of the art. In *Interactive Mobile Communication Technologies and Learning (IMCL), 2015 International Conference on*, pages 283–288. IEEE, 2015.

[21] J. J. LaViola, Jr. A discussion of cybersickness in virtual environments. *SIGCHI Bull.*, 32(1):47–56, Jan. 2000.

[22] L. Liao, Y. Ma, X. He, R. Hong, and T.-S. Chua. Knowledge-aware multimodal dialogue systems. pages 801–809. ACM, 10 2018.

[23] J. Lucas. Immersive VR in the construction classroom to increase student understanding of sequence, assembly, and space of wood frame construction. *Journal of Information Technology in Construction (ITcon)*, 23(9):179–194, 2018.

[24] J. Mathews. *Development and evaluation of spherical microphone array-enabled systems for immersive multi-user environments*. PhD thesis, Rensselaer Polytechnic Institute (RPI), Troy, NY, 2021.

[25] J. Mathews and J. Braasch. Sparse iterative beamforming using spherical microphone arrays for low-latency direction of arrival estimation in reverberant environments. *Journal of the Audio Engineering Society*, 69(12):967–977, 2021.

[26] M. Murray. Realizing the vision of a vr-based learning environment. *Medicine Meets Virtual Reality: The Convergence of Physical & Informational Technologies: Options for a New Era in Healthcare*, 62:134, 1999.

[27] G. Sharma. Spatially aware interactions in large scale immersive environments, 2019.

[28] G. Sharma, J. Braasch, and R. Radke. Interactions in a human-scale immersive environment: The CRAIVE-Lab. In *Cross-Surface 2016, in conjunction with the ACM International Conference on Interactive Surfaces and Spaces*, 2017.

[29] G. Sharma, J. Braasch, and R. J. Radke. Interactions in a human-scale immersive environment: The craive-lab. In *Cross-Surface 2016, in conjunction with the ACM International Conference on Interactive Surfaces and Spaces*, 2017.

[30] A. G. Solimini. Are there side effects to watching 3d movies? a prospective crossover observational study on visually induced motion sickness. *PloS one*, 8(2):e56160, 2013.

[31] Y. Song, R. Yan, X. Li, D. Zhao, and M. Zhang. Two are better than one: An ensemble of retrieval- and generation-based dialog systems. 10 2016.

[32] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J. Y. Nie, J. Gao, and B. Dolan. A neural network approach to context-sensitive generation of conversational responses. *NAACL HLT 2015 - 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pages 196–205, 6 2015.

[33] Z. Wu. Multi-object tracking and association with a camera network, 2014.

[34] P. Ziegler, D. Roth, A. Knots, M. Kreuzer, and S. von Mammen. Simulator sick but still immersed: A comparison of head-object collision handling and their impact on fun, immersion, and simulator sickness. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 743–744. IEEE, 2018.