Voice Interaction for Augmented Reality Navigation Interfaces with Natural Language Understanding

Junhong Zhao[1], Christopher James Parry[1], Rafael dos Anjos[1], Craig Anslow[2], and Taehyun Rhee[1]

[1]Computational Media Innovation Centre (CMIC),Victoria University of Wellington
[2]School of Engineering and Computer Science,Victoria University of Wellington

***Abstract*—Voice interaction with natural language understanding (NLU) has been extensively explored in desktop computers, handheld devices, and human-robot interaction. However, there is limited research into voice interaction with NLU in augmented reality (AR). There are benefits of using voice interaction in AR, such as high naturalness and being hands-free. In this project, we introduce VOARLA, an NLU-powered AR voice interface, which navigate courier driver delivery a package. A user study was completed to evaluate VOARLA against an AR voice interface without NLU to investigate the effectiveness of NLU in the navigation interface in AR. We evaluated from three aspects: accuracy, productivity, and commands learning curve. Results found that using NLU in AR increases the accuracy of the interface by 15%. However, higher accuracy did not correlate to an increase in productivity. Results suggest that NLU helped users remember the commands on the first run when they were unfamiliar with the system. This suggests that using NLU in an AR hands-free application can make the learning curve easier for new users.**

***Index Terms*—Augmented Reality, speech recognition, natural language understanding (NLU), speech interaction, artificial intelligence, intelligent interface**

I. Introduction

Augmented Reality (AR) is the blending of virtual objects superimposed into the real world [1]. AR is a growing market that has gained popularity in the past few years, enabling new and thrilling experiences on several different fields. Conventional interfaces for AR are motion controllers and mid-air hand gestures [2]. However, these methods occupy people's hands or body parts during interaction, and are not friendly to people with physical limitations or disabilities. New and more effective user interfaces are required to reach a wider AR user base.

Speech interaction opens up possibilities to interact with AR scenes in novel ways. It can be used to directly manipulate the virtual object, to conduct system navigation (mitigating repeated menu selection) and to collaborate with other touch-based interaction modes, extending the interactivity. Such additions will allow more flexible and natural interaction.

There are two main styles in speech interface in AR: *voice command style* in which users say pre-defined keywords, and *fluid conversation style* where users say spontaneous sentences to the AR system or avatar. Voice commands are the most conventional way for controlling purposes and has been widely used in many AR and virtual reality (VR) applications. Recently, speech recognition techniques have been significantly improved thanks to deep learning and big data [3]. These have brought speech interfaces to an unprecedented level of quality.

When integrating speech interfaces into an applications, speech recognition accuracy is a challenge [4]. The user's accent, various voice prints, and emotions all influence speech recognition performance. Low accuracy will reduce interaction efficiency and experience since the user has to repeat themselves from time to time.

Grammar or semantic understanding is another challenge in speech interfaces. Insufficient grammar comprehension often mitigates interaction flexibility. For example, without semantic understanding, systems cannot interpret “Hello" and “Hi" as having the same meaning or intention of greeting. In this case, AR designers have to explicitly define many possible commands to cover most real cases to enhance usability, which costs tedious labor. Also, it reduces the recognition accuracy since it raises the chance of commands overlapping with each other. Besides, when using voice interfaces in immersive AR scenes, a user can easily forget the scripted commands [5], [6]. As observed by Pascoal et al. [6], most users are unwilling to or struggle to remember rigid phrases or words. This becomes worse if the command list is very long. If the voice interface has semantic understanding, users have flexible ways of saying commands, and the cognitive barrier will be reduced.

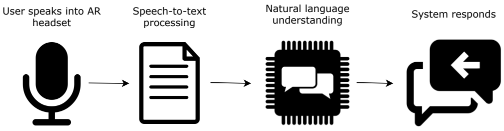

User speaks into AR headset
Speech-to-text processing
Natural language understanding
System responds

Fig. 1. High-level architecture of speech recognition engine

In this paper we study voice interfaces for AR applications that enable hands-free interaction. The main contribution is exploring NLU benefits in speech interaction in AR explicitly. A deep learning based natural language analysis component is overlaid to the speech recognition engine. The architecture is shown in Fig. 1. With the help of the NLU component, the

This project was funded by the Smart Ideas Endeavour Fund from MBIE and in part by the Entrepreneurial University Programme from TEC in New Zealand

978-1-7281-8579-8/20/$31.00 ©2020 IEEE

voice interface will not only know what the speaker said but also know their intention at a semantic level. The goal of this study is to investigate how effective is voice interaction with NLU for AR. The following research questions (RQ) will be studied:

- **RQ 1. Does natural language understanding increase voice interface accuracy?**

Implementing NLU offers more flexibility in the voice interface due to a larger grammar. We investigate to what degree the interaction accuracy of using speech in AR can be improved by such flexibility.

- **RQ 2. Does natural language understanding reduce task completion time?**

Will the more diverse and flexible grammar of NLU result in an reduced task completion time? Does an improvement in technology directly correlate to an increase in productivity?

- **RQ 3. Does natural language understanding help users remember voice commands?**

The NLU component identifies the intent of users voice commands. Users are able to say variations of the same command without pre-defined restrictions. By offering the users more ways to express an intent, do they may remember the command more easily?

To answer these questions, a specific prototype system called VOARLA (Voice Operated Augmented Reality Logistics Application) was designed. The task is to assist the user to locate packages and deliver them to the correct drop off location. The system keeps the users hands free of physical interaction with the package, while simultaneously manipulating the system state through voice. VOARLA features an interface that uses natural language understanding (NLU) in conjunction with speech-to-text. A user study was completed on VOARLA to compare with the voice interface without NLU.

II. Related Work

A. Speech Recognition

Speech recognition is the process of converting audio into text [7]. Employing speech recognition is proven to increase user experience in scenarios like car navigation [8]. Most speech recognition engines are based on deep learning algorithms such as Long Short Term Memory (LSTM) Networks [9], Time Delayed Neural Network (TDNN) [10] or end-to-end deep learning schemes based on Transformer [3], [11] to build up robust recognition models from huge datasets. Many online speech recognition engines have been developed and provide excellent services like IBM Watson [12], Dragon Naturally Speaking by Nuance [13], and Windows Speech Recognition [14].

B. Natural Language Understanding

The goal of natural language understanding (NLU) is to automatically extract intents, entities, keywords and semantic interpretation from text [15], [16]. NLU is often deployed in conjunction with speech recognition to analyze the text recognized from speech. Most NLU algorithms are based on deep learning methods like the attention-based encoder-decoder [17] and the recursive autoencoder [18]. NLU remains an active research area. Along with speech recognition engines, there are several natural language understanding tookits and accessible services, such as Watson Assistant [19] and Microsoft LUIS, developed by many advanced technology corporations. Normally, when creating domain specific applications with NLU, the possible intents must be known before deploying the assistant into the system.

C. Speech Interface for AR

Zhou, Feng et al [20] reviews ten years (1998 -2008) of tracking, interaction and display technology in AR. They conclude that creating appropriate interaction techniques for AR applications, allowing end users to interact with virtual content in an intuitive way, is very important. The overall goal of these interaction techniques is to enable manipulating AR content to be as easy as interacting with real objects.

Clark, Leigh et al. [21] surveyed 68 research papers to identify the trends, themes, findings and methods of empirical research on speech interfaces in HCI. Analysis of the research found that speech HCI work focuses on nine key topics, including modality comparison and how user memory affects speech interface interaction. Three papers assessed the effects of interface design on user memory. User recall of menu options in voice recognition systems was significantly impaired when five or more options were presented. Another study reported that more content was recalled when information was provided by a human speaker rather than a machine.

Chan, Zhen Yue et al. [22] proposed a methodology and design of a voice-controlled environment with an emphasis on speech recognition and voice control. The main components are Amazon Alexa and Raspberry Pi. Office users can easily control their office appliances with voice commands. They identified that the proposed system saves time and brings convenience to people. This study shows the value of integrating voice interactions into everyday life.

Irawati, Sylvia et al. [23] uses speech to interact with an AR scene. The results showed combining speech and paddle gestures improved performance in arranging virtual objects over using paddle input alone. The study identified that speech is suitable for control tasks, while gestures are suitable for spatial input such as direct interaction with the virtual objects. This example shows the benefit of applying speech as an interaction method in an AR scene.

There is limited research in NLU in AR speech interfaces. Two relevant research papers are the works from Maciej Majewski et al. [24], [25]. The paper presents the concept and implementation of an interactive system for controlling loader cranes. These works focus on improving voice interaction effectiveness between users and system actions when controlling modern machines in conditions of difficulty or increased risk. A semantic understanding module are delicately designed using binary neural networks. This work, however, only focusses on the NLU incorporation strategy. NLU effectiveness is not well studied and evaluated.

III. VOARLA: System Design

A. Overview

To well study the research questions that have been proposed in Section I, we design VOARLA[1] (Voice Operated Augmented Reality Logistics Application) as our prototype system. VOARLA is an AR system designed to manage logistics jobs and direct courier driver while making package deliveries. The system is hands-free as the user have to occupy their hands for courier holding in the task therefore need other interface channel to conduct informative navigation interaction. Here we choose to let the user conduct such interacts with voice only. The verbal interaction will guide users step by step on how to complete the tasks according to different physical conditions. Fig. 2 shows the system directing the user back to the start, avoiding obstacles along the way. The system features a range of voice interactions, such as offering and accepting jobs. Once chosen, the delivery path will be shown with virtual arrows.

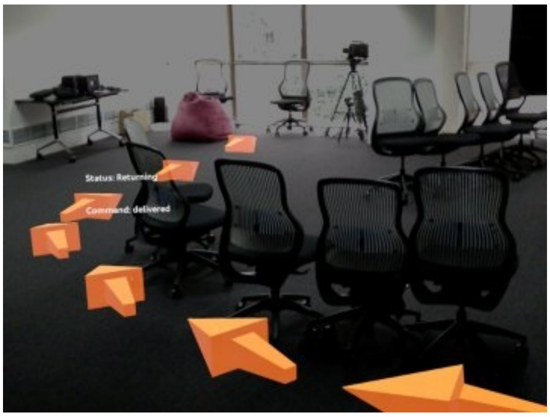


Fig. 2. VOARLA uses virtual arrows to direct user around obstacles

The user completes jobs by delivering packages between the locations, under system direction. When a job is accepted, the system uses virtual arrows and voice commands to direct the user to the pick-up location. Once the user confirms the package is received, the system directs the user to the drop off location. The user must confirm drop off before returning to the start location to await the next delivery job. A state-machine architecture is used for simplicity of all system states. The top left corner of the user's field-of-view (FOV) displays the current system state, the previous spoken command, and whether the engine is processing a voice command.

B. System Features

We develop VOARLA to have the following features to make sure its practicability as a completed AR system and its interactivity that powered by the voice technology.

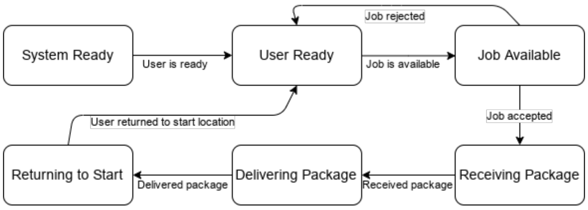


Fig. 3. System state-machine architecture

1) Flexible spoken commands: VOARLA features a NLU-powered voice interface with nine primary user commands as the semantic categories or intentions like "accept job", "package received". All the user allow diverse speaking way of their commands. As long as they have the same intention, they will perform the same action.

2) Highlight path, package and drop-off location: Bright orange virtual floating arrows highlight the best path to the target location, the correct package, and the specific drop-off location. The arrows are designed to be highly noticeable and direct the user between locations effectively.

3) Redirect: This feature redirects the user to an alternative target location. In the real world, there could be an unforeseen event which blocks the highlighted path. If the user finds the highlighted path blocked, they can inform the system path blocked by voice. The system responds with a voice clip confirming a successful redirect request. The target location is then changed to the next suitable location. This feature is to simulate a real world scenario in which an unforeseen event occurs.

4) Distance check for robust interaction: VOARLA contains a distance check feature to prevent errors. When the user informs the system they have received or delivered a package, the system checks the distance between the current target location and the user. This is to prevent the system state being able to progress further than it should. For example, if the user was at the start location in the receiving package state, the target location would be the package pick up location. If, due to an error or the user speaking, the voice interface understood a "received package" intent, it should not change into delivering state as the package has not yet been received. The distance check prevents this.

5) System voice for feedback: A voice clip is played every time the system changes state to notify the user direction and gives updates like "System ready", "Find package", etc. They are synthesized with a female voice. The systems response to a user speaking can be playing voice clips and turning virtual arrows on/off. The response is dependent on the current state, location and target location.

6) Help: At any stage of system execution, the user can say "help" to get assistance. The system voice will repeat the previous instruction for a reminder. An example is "Find and receive the package". This helps the user remember what they should do. This help function is an error handling technique designed to reduce the need of another humans help. The

[1] VOARLA Video:https://www.youtube.com/watch?v=gocOzvZTM9w&t=41s

help function also reduces the risk of a user removing the AR headset mid-test.

C. Multiple Delivery Simulations

Multiple package delivery simulations was built for the user evaluation to make sure enough interaction variations respect to different physical conditions. Three different package delivery jobs are setup for the user to complete and one exceptional job for the user to reject. There are five notable locations and virtual arrows directing the user between them. The start location and package pick up location is consistent throughout the test. The drop off location for the packages changes with each delivery to simulate the real world. Fig. 4 provides a birds-eye view of the deliveries.

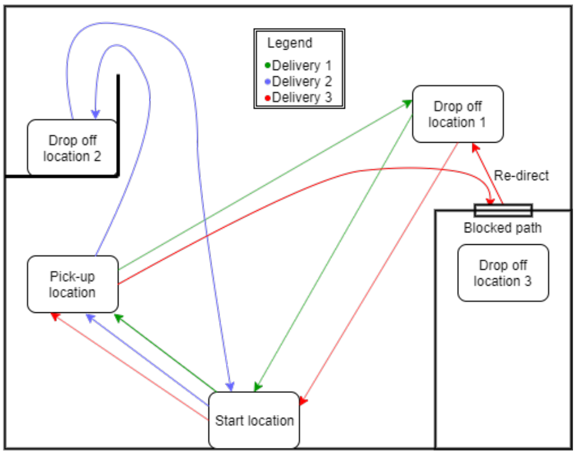


Fig. 4. Birds-eye view of delivery simulations

The first package delivery is between three locations in a triangle shape in the middle of the room. This job was deliberately made simple to help the user become familiar with the technology. The second and third deliveries each introduce an extra level of complexity. Job two involves delivering the package to a location out of sight, behind a wall. The virtual arrows are not visible through the screen, further simulating the real world. The user must follow the systems directions around a corner to find the correct drop off location. The third job directs the user through a blocked door. When the user finds their path blocked, they must notify the system by expressing the intention “path blocked". The system then reroutes the user to an alternative drop off location.

Three voice interactions were vital to completing the test. These were “help", “reject job" and “path blocked". “Help" could be used at any time to remind the user what they were supposed to do. “Reject job" is intended for use when the system offers the user an unwanted job. If the user accepts an unwanted job, the system replies “Unauthorized" and continues on to offer the next available job. “Path blocked" is used for when the path is blocked and the user needs the headset to reroute them. This is simulating an unforeseen obstacle.

D. Implementation

VOARLA was developed on the Magic Leap One AR head-mounted display (HMD). The Magic Leap runs the Lumin OS and provides Magic Leap Remote, a program that simulates the various Lumin SDK API to run Magic Leap applications locally. We develop all the prototype system in Magic Leap Remote PC simulator and then deploy into device. There are 4 jobs that come loaded on VOARLA, designed for the user evaluation.

The Watson Speech-to-text (STT) service is used to conduct speech recognition. The user's voice is sent to the service and returned as text. This speech recognition engine is speaker independent with high accuracy. While one limitation is the language models available. There is no New Zealand language model. The United States en-US language model was chosen due to the similarities. Another limitation is the network requirement. An offline STT engine is preferable for lower connection demand. However, the higher accuracy of the online service was prioritised over reduced latency here. The text is then sent to the Watson Assistant to do semantic interpretion.

IV. Evaluation

A. User Study Design

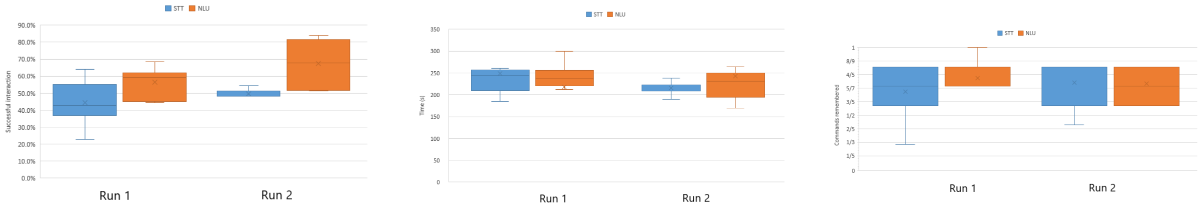
A user test evaluation was completed to investigate the advantages of using NLU in AR. The independent variable is the presence or not of NLU in VOARLA's voice interface. We used a within-subjects user study design, involving each participant engaging with both conditions. The order of prototypes was varied to reduce the learning effect bias.

1) RQ1-Accuracy: The voice interface accuracy is recorded for each test. This is calculated by the number of successful voice interactions divided by the total number of voice interactions attempted. The accuracy is recorded to show the advantage NLU offers over the same voice interface without NLU.

2) RQ2-Task completion time: The system contains a timer which activates once the user is ready and begins the test. The timer then finishes when the user returns to the start after the third and final delivery. This timer is to help us analyse whether the interface with NLU decreases the time taken to complete a hands-free task in AR over a STT-only interface. One consideration was the extra latency the NLU interface version requires. To offset this in the evaluation, the NLU latency was recorded each testing run and subtracted from the total run time.

3) RQ3-Words remembered: Participants are asked to try and use every voice command during each test run. The words remembered ratio was calculated through words remembered divided by total commands. “Yes" and “No" were not included in the words remembered calculation since we put them into the grammar for extra flexibility.

In the user study, participants were given an overview of the whole procedure and asked to read the information sheet and complete the study consent. Then they were asked to complete



(a) Voice interface accuracy (b) Task completion time (c) Voice commands remembered

Fig. 5. Results of user study on interface accuracy, task completion time and voice commands remembered

a basic user profile questionnaire about their AR experience and language skill (native or non-native speaker). After that, they were informed of the scene dynamics and the various possible voice interactions in the scene. They then enter into the first AR scene to start their test. After one test, they are asked to complete a interface evaluation questionnaire about their perceptions of task difficulties, system interaction Responsiveness and memory difficulties etc. Participants repeat the same procedures for the other tasks. Each run will cover all 4 jobs design in III-C and every participant will conduct two runs to evaluate their learning curves of command memory.

The participant was closely observed during the test. Data was recorded on how many voice interactions the user attempts and how many are successful. At the end of a test the number of total different interactions completed was recorded. The accuracy of the test subject's voice interactions was calculated by the total number of interactions divided by the number of invalid interactions. The number of possible voice interactions forgotten was recorded. This was determined by the total number of different possible interactions divided by the number of different interactions completed.

There were 15 participants who took part in the user study. The target age group was 18 years and up considering the intended use is in a logistics setting or work environment and people under 18 are less likely to be working there. The target participants required at least a basic understanding of computer systems and AR and have basic skills to operate the AR unit in general. Pilot studies were done before large-scale user study for assessing the feasibility of the full-scale study and developing and refine the research plan.

B. Results

1) RQ1-Accuracy: As seen in Fig. 5(a), the accuracy of the NLU-powered interface was consistently higher than the STT-only interface. The average accuracy of all STT-only runs was 46.6%. In contrast, the average accuracy of all the NLU-powered runs was 61.9%.

This shows that NLU offers higher accuracy than STT-only for a hands-free voice-operated AR application. The results yield another interesting trend. In the second run, once the user had some prior knowledge of the system, the maximum accuracy for NLU was far greater than the STT-only. This suggests that once the user has some prior knowledge, NLU is more effective than STT-only.

2) RQ2-Task completion time: As seen in Fig. 5(b), NLU had a slightly better average time taken for round one. While in round two the time taken became even worse. Results suggest that when the user had little knowledge of the system or possible commands, the presence of NLU helps users with the learning curve and can support voice interaction well in AR. Once the user had experience with the system, NLU was found not beneficial for completion.

While the introduction of NLU has been proven to increase the accuracy, and therefore decrease the number of invalid interactions. Minimizing the invalid interactions should save the user time while completing the task. However, this was not evident in the results. One potential explanation is the effect on the users' momentum. Processing a voice command in the NLU interface had a larger latency compared to the STT-only interface. When the user was waiting for the command to process in the NLU version, they lost their momentum which they may have otherwise kept in the STT version. This could have resulted in a longer time-taken for the NLU-powered interface, due to the loss of momentum.

3) RQ3-Commands remembered: The resulting number of commands remembered suggest NLU helped users remember the commands in the first run. There was a 10% advantage with NLU in first run. However there was little difference in the second run. This suggests NLU helps beginners with learning the possible voice interactions. Once the user has learnt the commands and surpassed the learning curve, NLU was not significantly better for remembering commands.

V. Summary

The contribution of this paper was to design VOARLA, a NLU-powered voice operated AR navigation application, to evaluate the NLU benefit in AR interaction. VOARLA is designed to navigate courier driver in locating and delivering packages. Voice interactions help the user find the fastest path between locations and identify packages. NLU module is investigated to address the rigid nature of a STT-only interfaces. NLU's benefit is evaluated in terms of accuracy, task completion time, and commands remembered.

The evaluation results show that the accuracy of our NLU-powered interface was consistently better than the STT-only interface. The average accuracy of the NLU-powered interface was 61.9%, compared to the 46.4% average accuracy of the STT-only interface. We would recommend using NLU in a hands-free AR voice navigation application if accuracy is a high priority. Besides, the NLU-powered interface offered a slightly better task completion time than the STT-only interface, and helped the user remember the commands. But it's only for first few performing of the task. So we would recommend to turn on NLU function for only the new users while keep other experienced user optional to this function.

It is not ideal that the system is restricted to one accent. Therefore, one of the future research could investigate solutions to support multiple accents, and potentially languages. The user could select the accent or language model prior to operating the system. Alternatively, the system could feature some form of accent detection at runtime, and then select the corresponding acoustic and language model. The challenges here would be getting all the various technologies working in unison.

Another possible future work could be directed at adding more complexity to the system. For example, the addition of tone analysis introduces an extra user interaction method. If users expressed a confused tone, systems using tone analysis could automatically offer help to the user.

References

[1] J. J. LaViola Jr, E. Kruijff, R. P. McMahan, D. Bowman, and I. P. Poupyrev, *3D user interfaces: theory and practice*. Addison-Wesley Professional, 2017.

[2] M. Billinghurst, H. Kato, and S. Myojin, "Advanced interaction techniques for augmented reality applications," in *International Conference on Virtual and Mixed Reality*. Springer, 2009, pp. 13–22.

[3] D. Wang, X. Wang, and S. Lv, "An overview of end-to-end automatic speech recognition," *Symmetry*, vol. 11, no. 8, p. 1018, 2019.

[4] S. McGlashan and T. Axling, "A speech interface to virtual environments," *Swedish Institute of Computer Science*, 1996.

[5] J. Lai and J. Vergo, "Medspeak: Report creation with continuous speech recognition," in *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, 1997, pp. 431–438.

[6] R. Pascoal, R. Ribeiro, F. Batista, and A. de Almeida, "Adapting speech recognition in augmented reality for mobile devices in outdoor environments," in *6th Symposium on Languages, Applications and Technologies (SLATE 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fur Informatik GmbH, Dagstuhl Publishing, 2017.

[7] D. Yu and L. Deng, *AUTOMATIC SPEECH RECOGNITION*. Springer, 2016.

[8] M. Westphal and A. Waibel, "Towards spontaneous speech recognition for on-board car navigation and information systems," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[9] A. Graves, N. Jaitly, and A.-r. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," in *2013 IEEE workshop on automatic speech recognition and understanding*. IEEE, 2013, pp. 273–278.

[10] R. Shaharin, U. K. Prodhan, and M. Rahman, "Performance study of TDNN training algorithm for speech recognition," *International Journal of Advanced Research in Computer Science & Technology*, vol. 2, no. 4, pp. 90–95, 2014.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[12] IBM, "IBM Watson," https://www.ibm.com/watson.

[13] Nuance, "Dragon – World's best-selling speech recognition software," https://www.nuance.com/en-nz/dragon.html.

[14] "Use voice recognition in Windows 10," https://support.microsoft.com/en-us/help/4027176/windows-10-use-voice-recognition.

[15] S. Trott, M. Eppe, and J. Feldman, "Recognizing intention from natural language: clarification dialog and construction grammar," in *Workshop on Communicating Intentions in Human-Robot Interaction*, 2016.

[16] M. Hamroun and M. S. Gouider, "A survey on intention analysis: successful approaches and open challenges," *JOURNAL OF INTELLIGENT INFORMATION SYSTEMS*, 2020.

[17] C. Li, Y. Du, and S. Wang, "Mining implicit intention using attention-based rnn encoder-decoder model," in *International Conference on Intelligent Computing*. Springer, 2017, pp. 413–424.

[18] T. Kato, A. Nagai, N. Noda, R. Sumitomo, J. Wu, and S. Yamamoto, "Utterance intent classification of a spoken dialogue system with efficiently untied recursive autoencoders," in *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, 2017, pp. 60–64.

[19] IBM, "Watson Assistant," Oct 2017, https://www.ibm.com/cloud/watson-assistant/.

[20] F. Zhou, H. B.-L. Duh, and M. Billinghurst, "Trends in augmented reality tracking, interaction and display: A review of ten years of ISMAR," in *2008 7th IEEE/ACM International Symposium on Mixed and Augmented Reality*. IEEE, 2008, pp. 193–202.

[21] L. Clark, P. Doyle, D. Garaialde, E. Gilmartin, S. Schlögl, J. Edlund, M. Aylett, J. Cabral, C. Munteanu, J. Edwards *et al.*, "The state of speech in HCI: Trends, Themes and Challenges," *Interacting with Computers*, vol. 31, no. 4, pp. 349–371, 2019.

[22] Z. Y. Chan and P. Shum, "Smart office: A voice-controlled workplace for everyone," in *Proceedings of the 2nd International Symposium on Computer Science and Intelligent Control*, 2018, pp. 1–5.

[23] S. Irawati, S. Green, M. Billinghurst, A. Duenser, and H. Ko, "An evaluation of an augmented reality multimodal interface using speech and paddle gestures," in *International Conference on Artificial Reality and Telexistence*. Springer, 2006, pp. 272–283.

[24] M. Majewski and W. Kacalak, "Conceptual design of innovative speech interfaces with augmented reality and interactive systems for controlling loader cranes," in *Artificial Intelligence Perspectives in Intelligent Systems*. Springer, 2016, pp. 237–247.

[25] M. Majewski and W. K., "Human-machine speech-based interfaces with augmented reality and interactive systems for controlling mobile cranes," in *International Conference on Interactive Collaborative Robotics*, 2016, pp. 89–98.