AR-TO-KID: A speech-enabled augmented reality to engage preschool children in pronunciation learning

IOP Conf. Series: Materials Science and Engineering **979** (2020) 012011 doi:10.1088/1757-899X/979/1/012011

AR-TO-KID: A speech-enabled augmented reality to engage preschool children in pronunciation learning

M Y F Aladin [1,2] *, A W Ismail [1,2], M S H Salam [1,2], R Kumoi [1,2] and A F Ali [1,2]

[1] ViCubeLab Research Group, Universiti Teknologi Malaysia, 81310 Johore, Malaysia
[2] School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia

* yahyfekri@gmail.com

Abstract. AR-TO-KID is an application produced for preschool children between ages five to six years old with an Augmented Reality (AR) application. The significant purpose of AR-TO-KID is to improve the pronunciation of the children in English. Hence, this paper discusses an AR application with speech input. The detection of the children speech input when they need to pronounce the words correctly, and they need to have critical thinking to identify the environment suit with the 3D objects that they will utter the word. Educational technology should be interactive and attractive for 5 to 6 years old preschool children learning; however, some at preschool teachers still used the conventional methods in teaching and children are not fully engaged with the method. Therefore, this project is to design and develop an interactive AR tool called AR-TO-KID for preschool children in pronunciation learning and teaching. This paper presents the evaluation and testing for preschool children with non-native English speaking. The article ends with results and discussion.

1. Introduction

In Malaysia, speaking in English as a second language is an excellent way to get a grasp of the language. We believe that speech is good at traversing interaction barrier, especially to young children because it lets them cut through reading difficulties and diffidence with a natural or more familiar way of communicating in their age [1]. However, minimal use of the language at home or in school has made non-native young children feel less motivated to use the language [2]. Speech recognition can be useful for the purpose of language learning, where it can teach proper pronunciation and help young language learners develop fluency with their speaking skills. This paper explores how Augmented Reality (AR) and speech recognition technologies could be used together for English pronunciation learning with children that are not native English speakers. We aim to examine the potential of AR to work with speech-enabled system for English pronunciation learning to pre-school children in Malaysia who do not speak English as their first language.

AR is a technology that required a tracking system to bring the virtual content appears on the top of the real-world [3]. Mobile or smartphones are capable to execute AR application, the camera can be a display device to capture the targeted features or known as a marker. The marker is a unique pattern that used to store or register the AR content. The AR tracking process is taking on a marker detection and a marker recognition [4]. The marker must to remain perceived by the AR application otherwise the AR

content will be disappeared [4]. User interaction using speech input is usually used with gestures in AR [5][6]. Speech recognition requires the list of grammar to be retrieved by the AR systems.

The proposed idea is the mobile AR application with speech-enabled system to be more significant to children 5-6 years old as it involves logical thinking for the kid to acknowledge the right and wrong based on the situation [7]. The concept of the real problem applies in AR to create a fun learning environment. Therefore, this paper will assess the effectiveness of using AR for English pronunciation learning, and it also discovers if combining speech input with AR cues facilitates the non-native English speaking children to learn English pronunciation. The main contributions of the paper include insights into how speech input with AR can enhance the children's interaction with an AR application and influence their learning experience. The more that is known about the effectiveness of speech input and AR for non-native young children's English language learning, the more effective learning approaches can be developed [8][9].

2. Methodology

The primary purpose of this application is the use of speech recognition with AR implementation. The proposed method has been described in the methodology as presented in Figure 1.

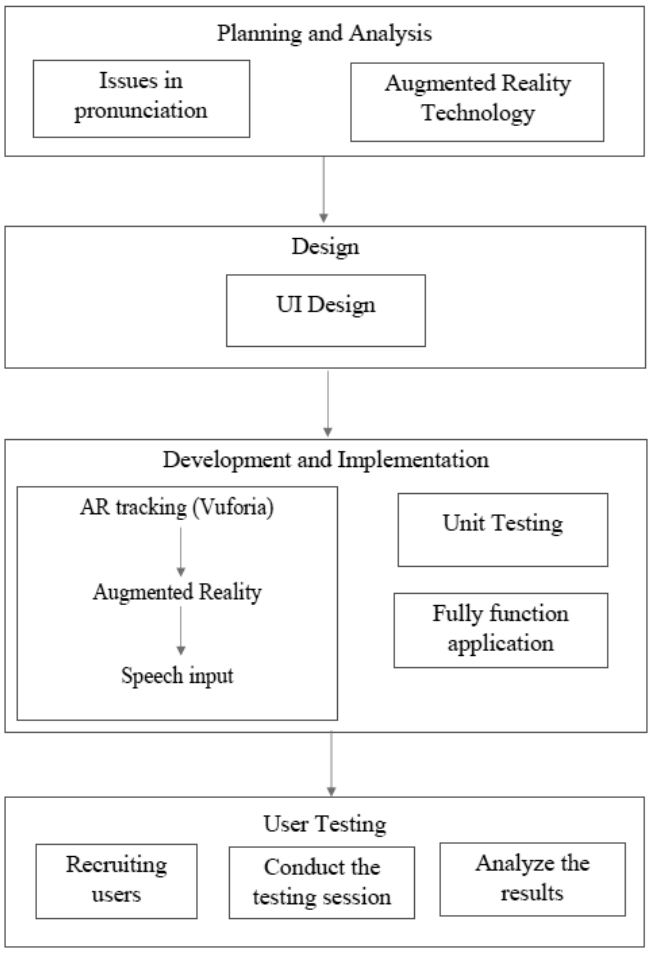


Figure 1. Research Methodology.

Figure 1 shows the methodology as a guideline to this development stages. There are four phases in this methodology which are Planning and Analysis phase, Design phase, Development and Implementation phase and User Testing phase. These phases are the phases that we followed in developing our application—visiting the preschool community to collect data and to know the issues

that may occur during the learning lesson. The issue that arises is the kids have problems with their pronunciation. After discussion, the solution for this issue is by making an AR application where the kids will pronounce the correct word that matches the object display to make the object (answer) appear. Through this project, we are using the Unity 2018 software to create the 3D object to apply the AR. Therefore, we designed the UI design using the same software. In the third phase, development is to enable AR by using the Vuforia, tracking for the application related to the AR and with the speech input. The fourth phase is about user testing of the application. In this phase, the application will be tested by the user for the evaluation of the application. Our users were preschool children with age between 5 to 6 years old.

2.1. Phase 1: Designing the User Interface (UI).

The usage of speech recognition is to focus on improving the pronunciation of the children as well as their logical thinking. The application requires the kid to speak the answer based on their logic when they understand the AR objects scene that appeared after the scanning on the marker. That is the reason why this application includes to improve their logical thinking. The right answer will pop out in a 3D shape after the user speaks the correct answer with the correct pronunciation. There will be three different 3D objects will show up on each scenes when the user plays the application.

In workspace setup as in Figure 2, the children were sitting on the chair while holding the smartphone and hover the camera to the marker. The marker is put on the table. When the children hover the camera to the marker, the 3D object will appear on the screen. Speech was recorded in classroom, the classroom environment sometimes produce unneccessary noise, and it has overlapped the speech of the children when they are speaking in the microphone. However, by using handheld device with the built-in microphone, the utterance more clear if the user sits close enough with the device.

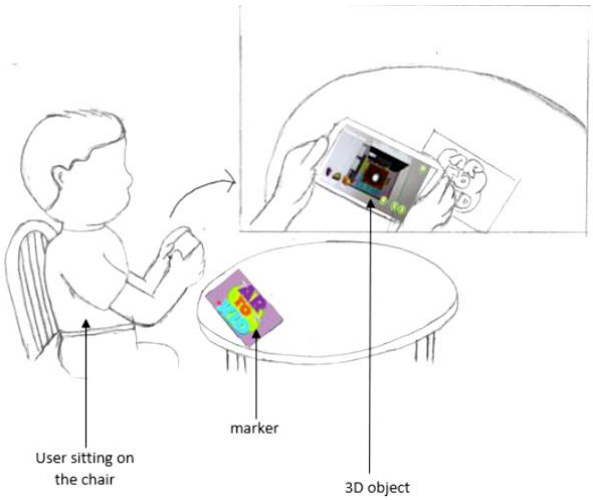


Figure 2. AR-TO-KIT Workspace.

Once the application starts, the camera with scan the marker and pop up question panel appear on the screen. The questions are different following each of the scene created for this application. The user can click on the 'X' button on the top right of the question panel once they have understood the question to close the panel (as in Figure 3 (a)). The grape, rock and vase images displayed on the screen are the options for the user to choose the answer. The user can hear the correct pronunciation of each of the solutions by clicking those images. The 'Speech Detection' on the top of the screen is for the user to know the words that the application detect. If the user taps on the 'Setting' button that located on the top right of the screen, it will display the setting panel, as shown in Figure 3 (b). There are three buttons on the panel, which are the 'Question' button, 'Instruction' button and 'Home' button. When the user

clicks on the 'Question' button, it will display back the question panel, as shown in Figure 3 (a). If the user hits the 'Instruction' button, it will display a pop-up instruction page, as shown in Figure 3 (c). When the user tap on the 'Home' button, it will return to display the home screen, as shown in Figure 3 (a). The user can tap on the 'X' button to close the setting panel (as in Figure 3 (b)).

Figure 3. User Interface Design.

Figure 3 (d) shows the scenes available in this application. The 'Coming Soon' button is to show the user that there will be more scenes coming soon. The number on each of the button shows the number of the scene provided in the application. When the user clicks on the first scene that indicates '1', it will display the first scene, as shown in Figure 4 (a). If the user clicks on the second scene that indicates '2', it will display the second scene, as discovered in Figure 4 (b) and for the third scene that indicates '3', the last scene will be displayed as shown in Figure 4 (c). If the user clicks on the 'Next' button, it will go back to the home screen, as shown in Figure 3 (d).

(a) Scene 1 (b) Scene 2 (c) Scene 3

Figure 4. 3D Scene appears on the marker once the scene is selected.

2.2. Phase 2: Enabling AR Tracking

AR is an interactive experience of a real-world environment where the objects that live in the real world are enhanced by computer-generated perceptual information, sometimes across multiple sensory modalities. In this project, we develop the interactive AR application named as AR-TO-KID by using Unity 3D version 2018 as a rendering engine with the Vuforia SDK for the tracking system to convert our marker into a feature.

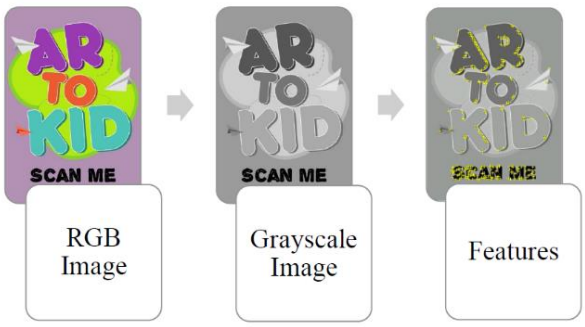


Figure 5. Creating a marker using Vuforia SDK.

Figure 5 shows our RGB marker has been converted into features, we used a single marker, coloured image in RGB as our image target and uploaded it into the database in the Vuforia SDK. This features-based tracking system will convert our image target from RGB to grayscale version to identify the features that can be used for recognition and tracking.

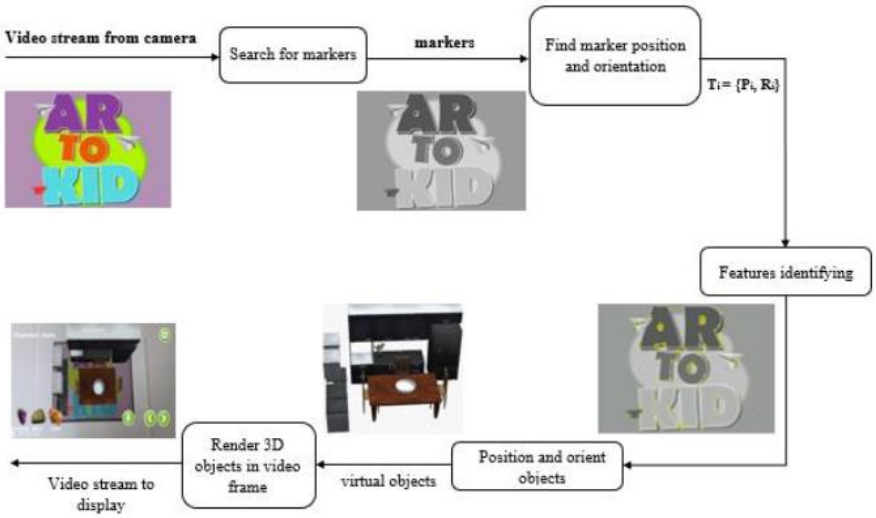


Figure 6. Features-Based Tracking Process for AR-TO-KID.

The recognition process is presented in the figure 6, the marker is defined as the image target. We improved the performance of our target by enhancing the visibility of these features through adjustments to our target's design, the rendering and scale of our target and how it is printed. Image target represents the image that the Vuforia Engine can detect and track. The image is converted to the binary image, as shown in Figure 6 and find the position and orientation of the image. The image positions and orientations relative to the camera are calculated using Ti = {Pi, Ri}. Then, the Engine detects and tracks the features that are naturally found in the image by comparing these natural features with a known target resource database. By using Ti, the 3D virtual objects are transformed to align them with the marker and rendered them in the video frame.

2.3. Phase 3: Speech-Enabled System with AR

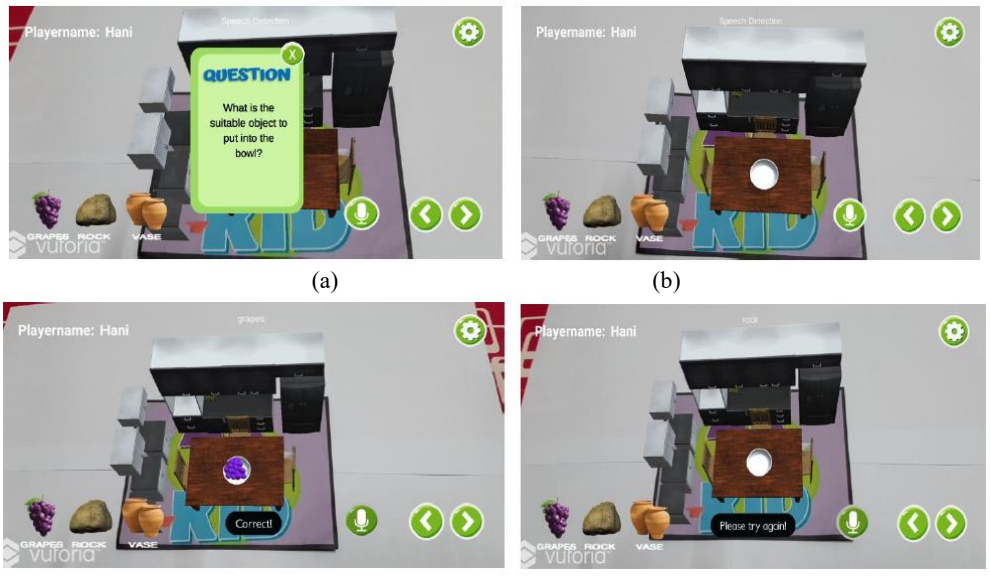
For speech input, we were using IBM Watson Speech to Text service on the IBM Cloud, which this service will convert the human voice into written text. It provides speech transcription capabilities for our application. It continuously updates and restores its dictation as it accepts more voice inputs.

In AR-TO-KID application, we have decided to use this service to make our AR object moves using speech. The list of grammars used in this system was:

a. Scene_1_Grammar: "grapes", "rock", "vase"
b. Scene_2_Grammar: "ball", "teddy bear", "television"
c. Scene_3_Grammar: "flower", "lollipop", "pencil"

Table 1. List of the English Pronunciation.

Scene and Questions	Pronunciation words
Scene 1: as in Figure 3 (a)	Grapes, Rock and Vase
Question: *What is the suitable object to place on the bowl?*	Answer: Grapes
Scene 2:as in Figure 3 (b)	Ball, Teddy Bear and Television
Question: *What is the suitable object to put on your bed?*	Answer: Teddy Bear
Scene 3: as in Figure 3 (c)	Flower, Lollipop and Pencil
Question: *What is the suitable object to put into the vase?*	Answer: Flower



(a) (b)

(c) (d)

Figure 7. AR-TO-KID works with speech.

The scene, as shown in Figure 7 (a) will appear the question panel, and the AR-TO-KID will play the wave to sound "What is the suitable object to place on the bowl?". After the children close the panel,

there are three images act as a 2D button that positioned linearly at the bottom left of the screen are the Grapes, Rock and Vase (as in Figure 7 (b)). These buttons represent the answer options for the user to choose from. The children can click on those buttons to hear the pronunciation of those particular icons. Then the children will need to say based on the selected object with the correct pronunciation, and the children have to touch and holds the 'Audio' button. The 3D object which is the grapes will fill the bowl, and a 'Correct' word will appear on the bottom of the screen once the children say the correct answer as shown in Figure 7 (c). If the children say the wrong answer or the other words besides the right answer, the screen will display 'Please try again!' statement, as shown in Figure 7 (d).

3. Results and Evaluation

As agreed by [12], there are no studies in the literature which define most appropriate usability guidelines for designing mobile-based AR applications for kindergarten kids. Age is a significant factor to be considered in developing applications for kids, especially when it comes to interaction techniques to be used in learning applications. The kids may not readily use and interact with the applications developed for adults. As a pre-test, the children were advised to fill out a short questionnaire before exploring the application. From the pre-test questionnaire, their background information including knowledge competency of English pronunciation were collected and we also get the teacher feedbacks. After that the post-test had carried out, they were asked the questions as listed in Table 2.

Table 2. List of Tasks in AR-TO-KID.

Tasks
1. User read the instructions by clicking the 'Instructions' button before the start.
2. User clicks the 'Credit' button.
3. User clicks the 'Start' button.
4. The user inserts their name before click 'Enter' button.
5. User clicks 'Enter' button.
6. The user hovers the camera to the marker.
7. User clicks on the 'X' button on a pop-up question after reading it.
8. User clicks on the image displayed on the screen to hear the pronunciation.
9. The user speaks the answer while click and hold the 'Audio' button.
10. User clicks on the 'Setting' button.
11. User clicks on the 'Question' button on setting panel to read the question again.
12. User clicks on the 'Instructions' button on setting panel to read the instructions again.
13. User clicks on the 'Home' button on setting panel.
14. User clicks on 'Next' button.
15. User clicks on the 'Back' button.
16. User clicks on the available scene '1' to '3' button on the 'Scene' page.
17. User clicks on the 'Next' button on 'Scene' page.
18. User clicks on 'Quit' button.

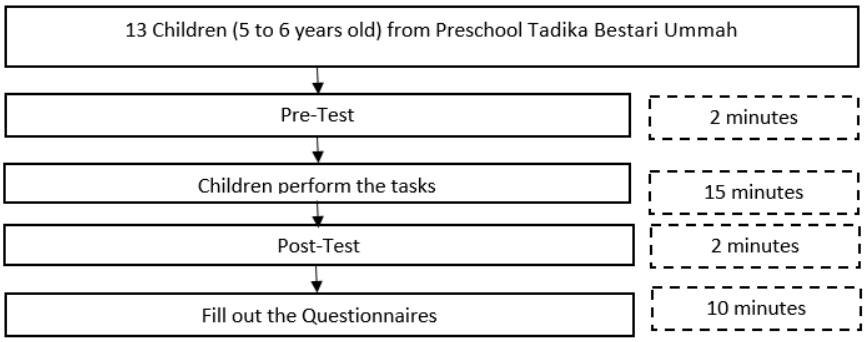


Figure 8. The experiment workflow.

At the end of the experiments, the children answered the questionnaire by using the smiley Likert scale (as in figure 8). Figure 8 shows the experiment workflow where the 13 children will run pre-test and post-test.

The learning session was conducted at the two kindergarten schools, their school will ensure that the children felt comfortable because they familiar with the environment, they will give more attention and cooperated well without feeling hesitant, this suggestion according to [11]. Thirteen children become our respondents. The learning session during the experiments has been recorded, the children hold the handheld device and hovers to the marker that found on the table, as shown in Figure 9.

Figure 9. Children play with AR-TO-KID application.

Figure 10. 7-Likert scale in AR-TO-KID for kids' usability ratings.

We used the smiley Likert scale for usability ratings, based on [13]. These emoji design used to replace the rank 1-7 Likert scale for children to understand the questions and straightforward answer the question. The scale in smiley style was redesigned to appear more colourful so it can suit the children to indicate their feeling. This list of questionnaires has been modified based on [13].

Table 3. List of the questionnaires [13].

#	Questionnaires
Q1	I enjoyed doing this activity very much
Q2	This activity was fun to do
Q3	I did not feel boring with this activity
Q4	This activity hold my attention at all
Q5	I would describe this activity as very interesting
Q6	I thought this activity was to improve my pronunciation
Q7	While I was doing this activity, I was thinking about how much I enjoyed it

Four children choose 5 and 7 scale for Q1. It is because they can pronounce the word better than others. The graph in Figure 11 shows that there are the children who have rated 3 and 4 where they did not enjoy the activity because the application has been interrupted by unstable internet. They think input voice is not really the best idea due to it required strong internet connection and also the application needs the quiet workspace. In Figure 12, two of the children choose to rate 5, 8 children have chosen 6, and 3 children have chosen 7. The children have fun with the application when they can see the 3D object appears on the top of the marker, and the application understand their voice command.

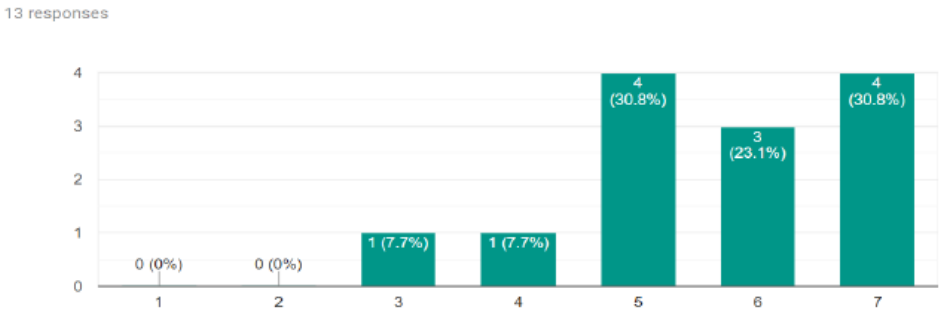


Figure 11. Response to the children enjoy this activity very much.

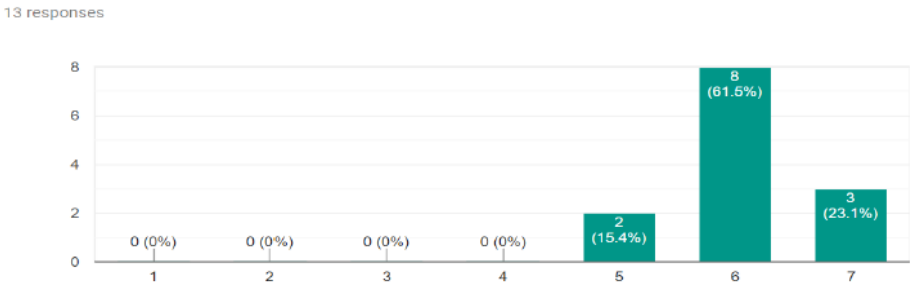


Figure 12. Response to the children found this activity is fun.

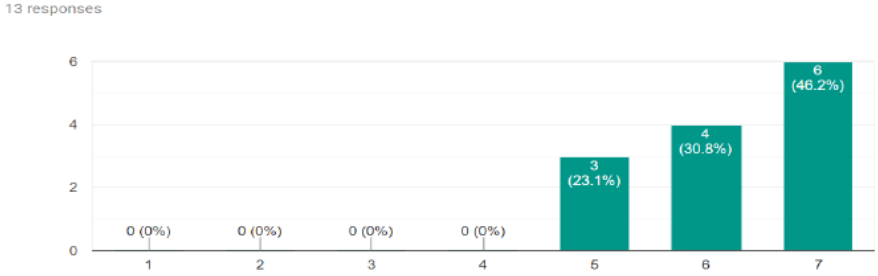


Figure 13. Response to the children found this activity is not boring.

Figure 13 has proven six children strongly agree where the application is not boring, while seven of them feel moderate. It is because the children understand the task they need to complete and they found AR is attractive and when the application brings out the voices for the question they have selected, it likes a virtual teacher has spoken to them.

Figure 14 shows how AR-TO-KID can get their attention. Only one participant has rated 5. Seven children have agreed the application hold their attention. Five children strongly agreed they can pay attention to the AR application.

Figure 15 children strongly agree where the application is impressive. The children think the application is interesting because they have to interact with the application using their voice. There is only one respondent who chooses four which is moderate. It is because the main task of the application, which is using input voice is quite hard due to demanding a strong internet connection and a quiet place. The application might not detect his/her voice during the testing session caused the user not really interested in the application.

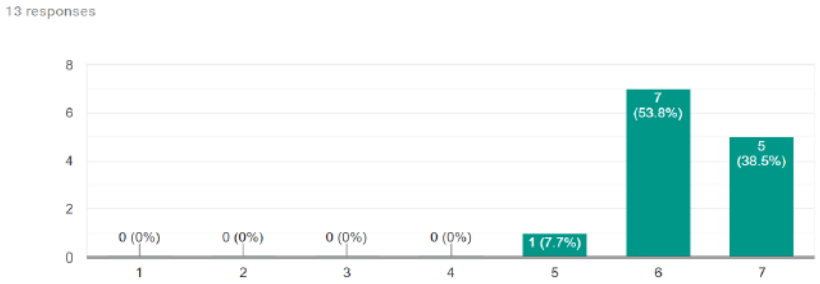


Figure 14. Response to the children found this activity hold their attention.

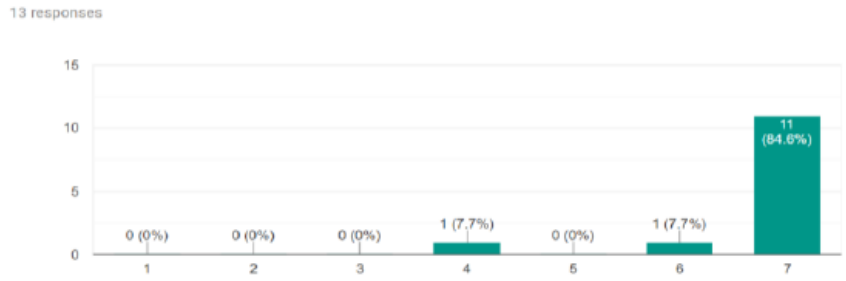


Figure 15. Response to the children found this activity very interesting.

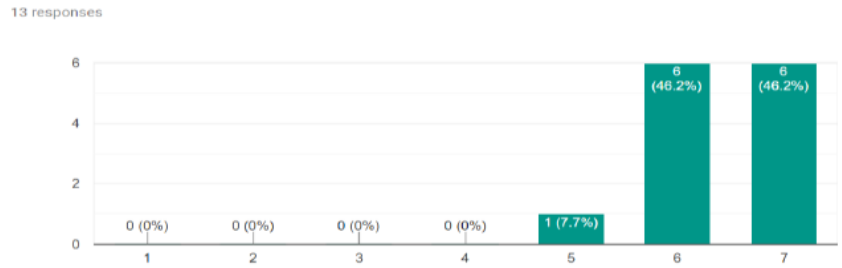


Figure 16. Response to the children found this activity improve their pronunciation in English.

Figure 16 shows the children strongly agree where the application helps to improve their pronunciation in English even they found it difficult to say it right because they are non-native English children. The children think the application has corrected their English when it comes to Scene 3. They have to interact with the application using their voice; they have to say the correct word to complete the

task. Only one participant feel disappointed with the application; the children have tried many times but still failed.

Based on the observation results, the participants who are the preschool children are non-native English. The application run the pronunciation must in the British accent. If the user says the word in an American accent or Malay accent, the application might have problems in detecting the user's pronunciation. The other issues arise; the application is an internet-based that require an internet connection to activate the speech recognition or otherwise, it will be not working. Therefore, a stable network connection is required during the testing. The feedback also, if the user says the word without holding the 'Speak' button, the application will not recognize the input, and it does not work. The environment with noisy too much voices will disturb the application. We need to ensure the application must be used in a quiet environment if they want to acquire high effectiveness of the application. If it is in the noisy surrounding, the application will have difficulty to detect the user's voice as there will be too many voice input in the background.

4. Conclusion

The primary purpose of this application is the use of speech recognition with AR implementation. The usage of speech recognition is to focus on improving the pronunciation of the children ages 5-6 years old as well as their logical thinking. The application requires the kid to speak the answer based on their logic when they understand the AR objects scene that appeared after the scanning on the marker. Marker detection is interrupting due to the light, condition of the classroom. The learning session was conducted at the two kindergarten schools, their school will ensure that the children felt comfortable because they familiar with the environment, they will give more attention and cooperated well without feeling hesitant. AR application using speech has improved preschool children pronunciation issues. The children can feel enjoy, and the AR application has improved the engagement. Children are easy to feel boring, AR is attractive and exciting for them, which could hold their attention.

Future works have been identified, and the speech recognition should be built-in with the application to avoid interruption due to the network coverage. The marker should be replaced with the more features extraction; it was suggested to replace the marker that contains more features to produce a robust marker recognition and overcome lighting disturbance [14]. Another suggestion, the AR-TO-KID application can use the marker-based interaction so children more feel realistic with the tangible interaction where they can hold the marker to interact [14]. The learning application also can be supported or assisted by invoke a conversational storytelling agent [15].

Acknowledgement

We would like to express our appreciation to Mixed and Virtual Reality Laboratory (mivielab), ViCubeLab at Universiti Teknologi Malaysia (UTM) for the facilities and technical supports. We also would like to thank to Pre-school Kindergarten, *Tadika Bestari Ummah* and *Tadika Al Ummah* which located in Johor Bahru for the user testing.

References

[1] Azman, H. (2016). Implementation and Challenges of English Language Education Reform in Malaysian Primary Schools. 3L: Language, Linguistics, Literature, 22, 65-78. https://doi.org/10.17576/3L-2016-2203-05

[2] Ansawi, B. (2017). Promoting the 3Es (Exposure, Experience, Engagement) in an English-Rich Rural Primary School Community. The English Teacher, 46, 30-42.

[3] Ismail, A., & Noh, Z. (2008). Augmented Reality Theory and Applications. M. Shahrizal, & M. Najib, Advances in Computer Graphics and Virtual Environment, 88-105.

[4] Ismail, A. W., Billinghurst, M., & Sunar, M. S. (2015, August). Vision-Based Technique and Issues for Multimodal Interaction in Augmented Reality. In Proceedings of the 8th International Symposium on Visual Information Communication and Interaction (pp. 75-82). ACM

[5] Ismail, A. W., & Sunar, M. S. (2015). Multimodal fusion: gesture and speech input in augmented reality environment. In Computational Intelligence in Information Systems (pp. 245-254). Springer, Cham.
[6] Aladin, M. Y. F., & Ismail, A. W. (2019, August). A review on multimodal interaction in Mixed Reality Environment. In IOP Conference Series: Materials Science and Engineering (Vol. 551, No. 1, p. 012049). IOP Publishing.
[7] Palupi, A. N. (2020). Use of Manipulative Media as A Stimulation of Ability to Understand The Concept of Early Children's Age. Early Childhood Research Journal (ECRJ), 2(1), 43-66.
[8] Potamianos, A. and Narayanan, S. (2003). Robust Recognition of Children's Speech. IEEE Transactions on SAP, 11(6):603–615, Nov
[9] Matassoni, Gretter, R., Falavigna, D., and Giuliani, D. (2018). Non-native children speech recognition through transfer learning. In Proc. of ICASSP.
[10] Eskenazi, M. (2009). An overview of spoken language technology for education. speech communication. Speech Communication, 51(10):2862–2873.
[11] Mispa, K., Mansor, E. I., Kamaruddin, A., & Hinds, J. (2016). Measuring usability and children's enjoyment of virtual toy in an imaginative play setting: A preliminary study. Journal of Theoretical and Applied Information Technology, 89(1), 45-52.
[12] Tuli, N., & Mantri, A. (2020). Usability Principles for Augmented Reality based Kindergarten Applications. Procedia Computer Science, 172, 679-687.
[13] Hall, L., Hume, C., & Tazzyman, S. (2016, June). Five degrees of happiness: Effective smiley face likert scales for evaluating with children. In Proceedings of the The 15th International Conference on Interaction Design and Children (pp. 311-321).
[14] Ismail, A. W. (2011). User Interaction Technique With 3D Object Manipulation in Augmented Reality Environment (Doctoral dissertation, Universiti Teknologi Malaysia).
[15] Ureta, J., Brito, C. I., Dy, J. B., Santos, K. A., Villaluna, W., & Ong, E. (2020, September). At Home with Alexa: A Tale of Two Conversational Agents. In International Conference on Text, Speech, and Dialogue (pp. 495-503). Springer, Cham.