CampusMaps: Voice Assisted Navigation using Alexa Skills

Revathi Vijayaraghavan, _, _, Dhiren Patel
Department of Computer Engineering & Information Technology
VJTI
Mumbai, India
{rvijayaraghavan, _, _}_b17@ce.vjti.ac.in, director@vjti.ac.in

***Abstract*— Voice assistants have become a common occurrence that are present in our smart-phones, speakers or dedicated devices like the Google Home and Alexa Echo. Over the years, these voice assistants have become more intelligent and the number of tasks that can be performed by them has increased. Of the many voice assistants that exist, Amazon's Alexa is one of the most compatible voice assistants, which can be programmed to suit specific use cases by the use of Amazon's Alexa Skills Kit. Through this paper, we leverage the power of Alexa's voice assistance by designing a navigation system for our college campus, which allows users to request directions in the most intuitive way possible. This is a cost-effective and scalable solution based on Amazon Web Services (AWS) Lambda.**

Keywords*— *Voice Assistant, Alexa Skill, Navigation System, Routing, Amazon Echo.

I. INTRODUCTION

Voice recognition and assistance has immensely increased in past years, and the use of voice assistants for day-to-day activities and basic tasks have become ubiquitous. Voice assistance has become popular because on an average, humans can type 40 words per minute, whereas they can speak upto 150 words per minute, making it the preferred form of communication. Voice assistants leverage this fact and offer a natural way of interaction, also providing easy access to those who are visually impaired. The most popular voice assistants in today's times are Google's Google Assistant, Amazon's Alexa, Apple's Siri and Microsoft's Cortana. As time has progressed, so have the capabilities of these assistants. Although these are the most popular assistants in the current times, voice recognition and assistance has existed since many years before. Over the years, voice recognition has evolved from being limited to a vocabulary of 1000 words as was the case with CMU's Harpy [1] in 1976 to being able to recognize full-fledged conversations almost like that of an adult human that we see today.

General tasks of voice assistants include setting reminders and alarms, giving weather updates, etc. However, their usage isn't limited to these tasks. They can also be employed for building systems that are tailored for specific purposes. Each of the voice assistants can be leveraged for their own strengths. Among the most popular ones, Amazon's Alexa is one of the most developer friendly and best in terms of compatibility among all voice assistants. [2] Amazon provides developers with the "Amazon Skills Kit" (ASK) [3] which can be used to build custom skills, which are like apps for Alexa, with self-defined functionalities. The ASK provides self-service APIs that can be used to build skills that range from games and music streaming to multimodal experiences and smart home control.

For this reason, we use Amazon Alexa for augmenting a skill to build a voice-assisted campus navigation system. This skill is based on the multimodal experience provided as part of the Amazon Skills Kit, that gives spoken navigation assistance as well as visual reference for viewing the path. We propose a methodology that has high accuracy while also being comparatively inexpensive. The built skill will be accessible on any device that can run Alexa, but the visual component of the skill is limited to Amazon's Echo devices that come with a built-in display. To thoroughly test our system, it has been deployed on an Amazon Echo Show 8.0 device. The Echo Show has an 8" display that can be used to display our multimodal response in the form of images. Any device that can run the Alexa voice assistant will be able to access our skill, and the voice assistance in the form of directions will remain the same.

The rest of the paper is organized as follows: Section 2 discusses background and basic terminologies. In Section 3, proposed system for Alexa based Navigation is discussed. Section 4 discusses the advantages of the system and how challenges in the system were addressed, and section 5 includes results. Conclusion and future scope are presented in Section 6 with references at the end.

II. BACKGROUND AND TERMINOLOGY

A voice assistant is a software agent that uses natural language techniques to interpret speech input given by users and responds via synthesized voice messages [4]. The first ever voice activated product was the Radio Rex, which responded to the word "rex", hearing which it would jump out of its house. This was formulated by the use of an electromagnet, as this was made in the year 1922, when modern computers did not exist. In the period from 1971 to 1976, DARPA funded R&D in Speech Understanding Research (SUR) program, the

outcome of which was CMU's Harpy, which could recognize over 1000 words. [1] In 1990, Dragon Dictate was rolled out to the public, as part of the DOS machine, which recognized speech using the Markov model. The limitations of the Dragon Dictate were addressed in Dragon NaturallySpeaking, which was released seven years later. In 1996, IBM launched MedSpeak, the first commercial product to recognize continuous speech. [5] The first modern voice assistants was Siri, released by Apple in the year 2011. Following this, Microsoft's Cortana and Amazon's Alexa were released in 2014, and Google's Google Assistant in the year 2016. [6] In 2017, Google Assistant had achieved a 95% accuracy in recognizing words in the English language.

A. Voice Assistants

1) Alexa: Amazon's Alexa is a voice assistant which runs on smart speakers (Echo devices) or smartphones. In addition to performing basic tasks like setting alarms/reminders, giving weather updates, playing music,etc, other features can be added to Alexa, called skills.

2) Google Assistant: This is a personal assistant by Google, which is available on smartphones and smart devices like the Google Home. In 2016, Google also launched Google actions, which is a developer platform to build apps for Google Assistant. [7]

3) Cortana: Cortana is a productivity assistant developed by Microsoft. Cortana uses the Bing search engine to answer questions asked by the user. Cortana's key features include reading an overview of mails and enabling voice-dictated replies, creating planners, recommending activities, etc [8].

4) Siri: Siri is the voice assistant developed by Apple. Siri was the first voice assistant to be rolled out to the public. A wide range of voice commands are supported by Siri, ranging from basic commands such as searching the internet and setting reminders to more complicated ones like engaging with third party apps on iOS. [9]

5) Comparison: All four voice assistants mentioned above have been popular and made life more convenient. However, these voice assistants are not perfect and have large scope for improvement. Siri was one of the first voice assistants in the market, yet it was considered inflexible and unnatural compared to other voice assistants. Cortana was removed from the iOS market in 2019 and from the Android market in 2020 [10]. With Google Assistant and Amazon Alexa as the two best choices for building custom applications, we compared the advantages and disadvantages of using either of them. Amazon has the feature of Alexa blueprints, which provides templates for getting started with skill development. Alexa Presentation Language (APL) provides many templates for visual responses as well. On the other hand, Google assistant lacks such control over the responses [11]. Further, Amazon also has easy-to-use APIs for integration with different applications and provides ease of building custom apps in the form of skills. A device with Alexa is claimed to have more compatibility with third-party applications and services compared to its counterparts [2]. The potential of Alexa and Echo devices to take on a range of different roles and functions in multi-user interactions makes it particularly relevant for our use case.

6) Amazon Echo and Echo Show: Amazon Echo, or Echo, are smart-speakers built by Amazon, that use the Alexa voice assistant. [12] The device can act as a simple voice assistant that does day to day tasks or also be combined with several smart devices like smart bulbs, smart fans and smart switches to name a few, thus facilitating home-automation. The different variants are the first-generation Echo, the Echo Dot, Amazon Tap, Echo Look, Echo Spot, Echo Plus, Echo Connect and Echo Flex. An important Echo variant that has been the primary device from the point of view of this paper is the Echo Show, which comes with an in-built display that aids multimodal responses. All of these devices vary in terms of size and presence of an inbuilt speaker, however the voice recognition software running on each is the same.

B. Navigation Systems

The purpose of navigation systems is to give directions to the user for the requested destination from a source location. Before the advent of GPS and navigation systems like Google Maps, navigation would be done using paper-printed maps. The invention of digital navigation was a game-changer providing dynamic routes and assistance. Mainly, directions can be provided by visual, audio and/or haptic means [13]. Speech-based navigation systems can fully tap the potential of voice assistant devices. Moreover, generating an audio output has less processing overhead resulting in quicker response time. However, for longer and complicated responses such systems can be supplemented by a visual component. Visual aids like maps help the user absorb complicated direction information in a relatively easier way. Haptic-based interfaces use the user's sense of touch to interact and give directions.

Outdoor navigation systems can use GPS (Global Positioning System) for positioning. However, radio signals cannot penetrate solid walls rendering GPS ineffective for indoor positioning [14]. In such cases, additional installations like beacons, sensor networks, etc are required which would increase the system cost by a fair amount.

C. Terminology

Some of the terminologies associated with Amazon Alexa that have been used in the paper have been elaborated upon in this section.

1) Intent: An intent is a representation of an action that responds to a user's spoken request. Intents may have slots (placeholders for arguments) [15]. In our case we have defined 5 Custom Intents which are as follows:

- DirectionIntent
- SourceIntent
- GreetIntent
- WhereAmI Intent

- WashroomIntent

We have also leveraged some of the built-in intents provided by Alexa Skills Kit like YesIntent, NoIntent, HelpIntent, etc.

2) Utterance: Utterances are phrases which will be most likely used for a particular intent [15]. Utterance sets should be unambiguous and should contain as many relevant phrases as possible. For example, in our use-case, "DirectionIntent" will have an utterance set consisting of phrases like: where is <location>, where can I find <location>, how do I reach <location>, etc.

3) Custom Slot Types: Custom slot type is a collection of possible values for a slot type. Custom slot types are used to represent items that are not included in Amazon's built-in slot types [15], which in our case, are names of locations in the campus.

4) Multimodal response: Multimodal is the addition of other forms of communication, such as visual aids, to the voice experience that Alexa already provides. Multimodal skills [15] can provide more information through visual aids leading to a better user experience. For any given route, it is difficult for the user to visualize and remember the entire path given by Alexa voice output. Voice response supplemented by a visual map component will help users find their way in the campus effortlessly. Our skill uses Alexa Presentation Language (APL) to render visual aids with every direction response.

5) Endpoints and Lambda Function: The endpoint for skill is the service to which Alexa sends requests when users invoke it. A custom skill can be hosted using AWS Lambda or as a web-service. We are hosting our skill's backend service as an AWS Lambda function [15].

III. PROPOSED SYSTEM: CAMPUSMAPS

We propose a voice assisted solution along with an addition of visual representation of the path to enhance user experience. The visual maps are also integrated with alexa skills, but will be visible only on alexa devices that have an inbuilt display. Smartphones that have the Alexa app can access the skill, but will not be able to render the visual component of the skill. The overall workflow is shown in Fig. 1. Once the skill has been invoked using the correct invocation *"open campus maps"* the user can ask for directions to a specific location. If the source location (location at which the echo device/location of user) has not been set, the skill will prompt the user to set a source location. Once this has been set, the user can ask for directions and the skill will compute the path and directions at the backend and return a set of directions along with a visual path in the form of an image to the user. We have additionally included features like finding the closest washroom or staircase. The user can also enquire about his own location by asking *"Where am I"* and this will return the preset source location. The built skill uses two languages that are supported by Alexa, namely English (EN) and Hindi.

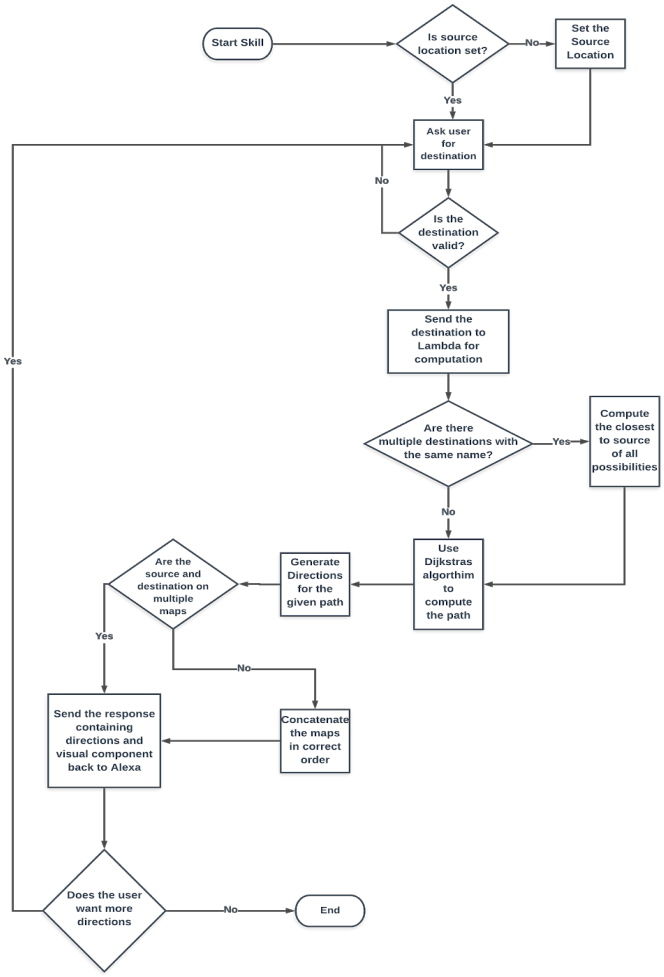

Fig. 1. Workflow of the system

A. Architecture

The user interaction with the system will take place through an Alexa enabled device (Amazon Echo Show). The user can request directions for a location within the campus and this location will be passed as input for the Custom Alexa Skill backend. The skill (backend) then processes this input and generates a path from a pre-set source location to the destination as mentioned by the user along with a map representing the route for the same. The generated directions and image of the map are returned to the alexa device, which then displays the visuals while speaking out directions to reach the given destination. Fig. 2. shows user interaction with the custom skill and how it is processed.

B. Modelling Maps

The campus has been divided into 3 regions. Each region has maps representing different floors. Each important location on the campus is stored as a node. The nodes are represented as pixel coordinates of each map image. A mapping of the node index to the node metadata is created, so that when we receive a location as input we can derive the pertinent information like the coordinates, map number and floor that the node is placed in. Connecting undirected edges

represent which two locations on campus are connected. The nodes and edges are stored in a static file that contains the pixel coordinates of the image, where each image is of the resolution 600 x 600. Fig 3 and 4 show mapping of location name to node number and mapping of node number to metadata of the node respectively.

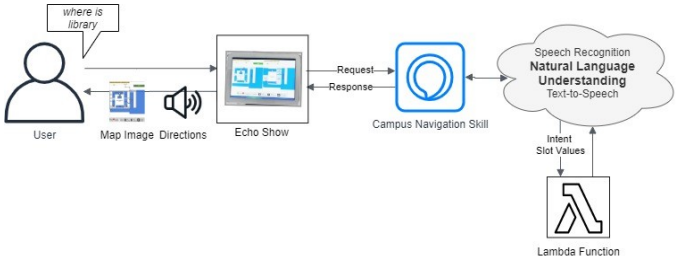

Fig. 2. User interaction with custom skill and how user requests are processed

C. Routing

This forms the crux of the system, and to provide the shortest and most efficient solution, we have used a modified version of Dijkstra Routing Algorithm. [16] In the modified version, we are not only finding the minimum distance (as is the case with traditional Dijkstras) but also the path from the source to destination. For special cases like washroom, our algorithm returns the closest washroom to the user.


```
{
  "boys hostel c": 2,
  "boys hostel d": 3,
  "cricket ground": 5,
  "girls hostel": 6,
  "football ground": 7,
  "mechanical gate": 16,
  "mechanical building entrance": 18,
  "xerox center": 22
}
```


Fig. 3. Mapping of node name to node number


```
"3": {
    "Node number": "3",
    "Node Name": "boys hostel c",
    "x_pos": "338",
    "y_pos": "338",
    "Type": "",
    "Floor": "0",
    "Building": "",
    "Map Number": "1",
    "Comments": ""
},
```


Fig. 4. Mapping of node number to node information

1) Extended routing to multiple maps: There are many use cases such that source and destination would be far apart (such as two ends of the college) and would thus not be in the same (distributed) map. For such use cases, the routing algorithm has been extended so that even if a user wants to go to a location which would not be in the same map the most efficient distance (and path) would still be found.

2) Floor Navigation: The algorithm supports navigation on different floors as well. This has been done by adding edges for every staircase present. In the event that there may be multiple staircases to reach a destination, the closest staircase will be chosen.

3) Directions: Providing the correct directions from a source to the destination is one of the most important parts of the CampusMaps. It is important to note that directions would vary from situation to situation- for instance someone entering from one end of the college would find some places to be on his/her right, but for another person exiting the college via a gate on the opposite end, the same place would be on his/her left. For this reason, it was essential to account for the direction the person was already walking in, and only then could the next direction be provided. For the special case where there is no previous direction to reference, i.e when the user is at the source location, a special provision has been provided such that the user is first told to turn in the correct direction after which the rest of the directions are provided. It has been ensured that the directions are provided in a language that sounds natural and conversational to the user.

D. Alexa Skill Development

Amazon Developer Console has been used to develop our Alexa skill. This platform is used to configure, build, test and maintain a skill after deployment. The initial configuration consists of setting an appropriate skill name, selecting languages for the skill (English-India and Hindi in our case) and choosing a model to add to the skill (Custom). The next step involves setting up build configurations which includes selecting an invocation name, defining intents along with their respective utterance sets, slots and defining an endpoint which will trigger actions every time the skill is invoked. Finally, the model is trained with the chosen configurations and built.

The endpoint for our skill is a lambda function. The core backend logic for the skill resides in this function. It is responsible for extracting intent, slot values and other metadata from the request object it receives every time the skill is invoked. A response object is constructed which has an appropriate speech output and a reprompt message (if needed). If the device invoking our skill has a screen, a map image is sent using Alexa Presentation Language [15] (APL). The images are pre-constructed to reduce response-time and are hosted on a separate server.

Users can interact with the system by giving voice commands. Amazon's Natural Language Understanding [15] (NLU) module is responsible for identifying utterances and their corresponding intents from the input and sending its findings to the skill's backend (lambda function) in the form of a JSON Object. The intents and slot values (if any) are extracted from this response object and an appropriate action is launched. For instance, if the intent identified is DirectionIntent with a slot value library, a function called direction() is called. This function will return text which gives directions to the library, which should be interpreted by the device as speech-output. Fig. 5 shows sample output map from Main gate (source location) to Canteen (destination). The

directions for the same case are as follows: *"Walk straight. Take the next left. Take the next right. Continue straight. Take the next right. You have arrived at the canteen. You have walked a total of 114meters"*

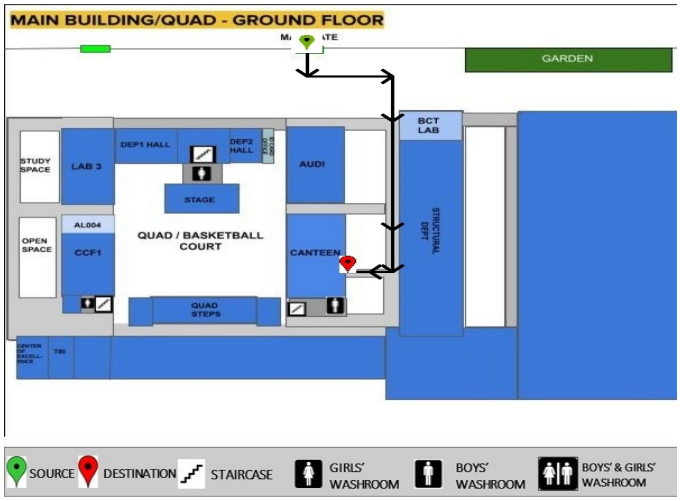

Fig. 5. Sample output image (Source: Main Gate, Destination: Canteen)

E. Use of Modular Programming

This system uses modular programming to ensure the functionalities of different components are separated from each other. Each sub-module focuses only one aspect of the desired functionality (for eg routing, visuals). This approach improves the maintainability and the readability of the code and is convenient to make any changes in future or to correct the errors. This is also advantageous because such a system can easily be applied to any other campus or establishment, by simply changing the underlying map, related nodes and slot values in the skill. The rest of the modules would require minimal changes to be applicable to the new campus.

IV. ADVANTAGES AND CHALLENGES

A. Advantages

1) Cost Effective: The system is built economically, compared to other navigation systems like indoor maps and GPS based systems, that require a huge budget. The only cost incurred in this system is that of the Echo device(s).

2) Scalable: With the use of Amazon Web Services Lambda and Google Drive to store the generated output of maps for each use case, the system can be scaled easily as the storage requirements are minimum.

3) Time required to build the system: Compared to other systems like indoor maps and GPS based systems, this system is far quicker to build. The map modeling for any campus can be done in a short period of time and since this is the only change required to extend this architecture, this system can be replicated for any other campus or establishment rapidly.

B. Challenges

The major challenge that we faced was the incorrect recognition of locations that were spoken by the user in both English as well as Hindi. This was resolved by taking the following steps:

1) Removing Utterance Conflicts: As our skill had custom intents, utterance conflicts between both custom-custom intents and custom and built-In utterances were encountered. This issue was resolved by making utterances in custom intents more descriptive to remove ambiguity. Redundant utterances were removed.

2) Testing intent resolution capability: As our skill's interaction model was being built, we used the utterance profiler [17], a feature in Amazon's Developer Console to test our model's intent resolution accuracy. Utterances to be tested were fed as input and the profiler identified corresponding intents and slots. Utterance set was modified whenever the test utterances did not resolve to the right intent followed by re-building the model.

3) Batch Testing using NLU tool: Natural Language Understanding [15] (NLU) evaluation tool in the developer console was used to test batches of sample utterances. The test set (aka an annotation set) consists of utterances mapped to the expected intents and slots. NLU evaluation score with the annotation determines how well the skill's model performs against our expectations [18]. The score was improved by removing ambiguities in the utterance set and adding synonyms for various slot values.

4) Testing Sample audio files using ASR Tool: The Automatic Speech Recognition [15] (ASR) Evaluation tool allows us to batch test audio files to measure the speech recognition accuracy of our skill. With this tool, we have tested and compared expected transcriptions against generated transcriptions for sample utterances. The issues uncovered during this test have been resolved by spelling out the words in utterance sets considering how they will be actually pronounced.

V. RESULTS

A. Unit Tests

The first step of testing was to test the individual components on our code editor itself. The following factors were considered to gauge correct working with each test case:

- Path is selected corrected
- Directions are given correctly
- Correct nodes are selected on the visual component
- Correct path is shown on the visual component
- All maps are shown in the case of source and destination being on different maps
- The maps are concatenated in the correct order before saving

Manual verification was done for each test case. All the test cases ran successfully.

B. Alexa Developer Console Test Simulator

Before deploying the skill to the echo devices, we ran the tests on the alexa developer console for testing skill in english and hindi. The output of the tests are shown in Fig. 6 and 7.

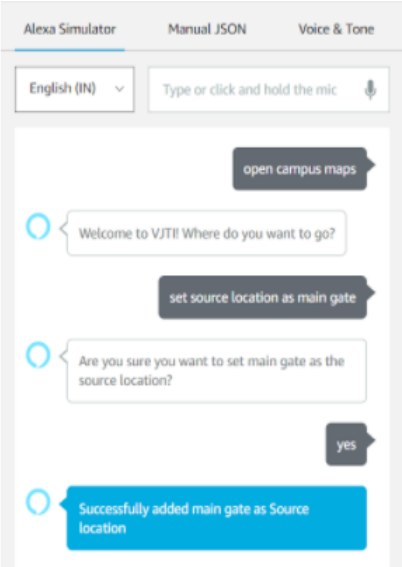

Fig. 6. Testing on console (English)

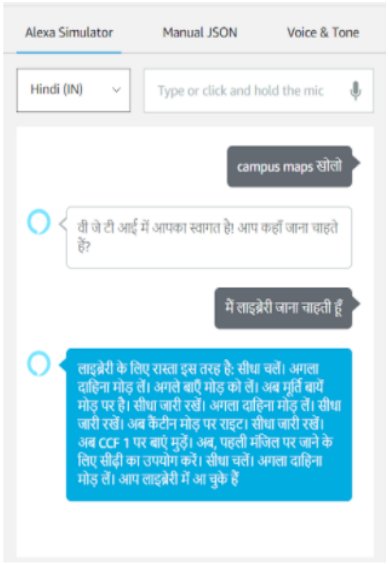

Fig. 7. Testing on console (Hindi)

VI. CONCLUSION AND FUTURE SCOPE

In this work, we have proposed a voice assisted navigation system for campuses and establishments where extensive systems like google maps or indoor maps aren't necessary or feasible. The purpose of the system was to leverage the power of Alexa Skills, which has been done efficiently as supported by the results. This prototype of the system makes the navigation within the campus an easier and more convenient task for new visitors. Currently, the position of the Alexa enabled device is static thus limiting the user to understand the directions in a single-go. As a future scope, this system can be further enhanced by the use of Beacons for positioning, which can provide higher accuracy and dynamic navigation.

REFERENCES

[1] Newell, Allen. "Harpy, production systems, and human cognition." Perception and production of fluent speech (1980): 289-380.

[2] O'Boyle, B. (2020, October 12). Google Assistant vs Alexa vs Siri: Battle of the personal assistants. Pocket-Lint, https://www.pocket-lint.com/smart-home/buyers-guides/124938-google-assistant-vs-alexa-vs-siri-personal-assistants, last accessed 2020/11/12.

[3] Build Skills with the Alexa Skills Kit | Alexa Skills Kit, https://developer.amazon.com/en-US/docs/alexa/ask-overviews/build-skills-with-the-alexa-skills-kit.html, last accessed 2020/11/12.

[4] Matthew B. Hoy (2018) Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants, Medical Reference Services Quarterly, 37:1, 81-88, DOI: 10.1080/02763869.2018.1404391

[5] Timeline of speech and voice recognition, https://en.wikipedia.org/wiki/Timeline_of_speech_and_voice_recognition, last accessed 2020/11/12.

[6] López G., Quesada L., Guerrero L.A. (2018) Alexa vs. Siri vs. Cortana vs. Google Assistant: A Comparison of Speech-Based Natural User Interfaces. In: Nunes I. (eds) Advances in Human Factors and Systems Interaction. AHFE 2017. Advances in Intelligent Systems and Computing, vol 592. Springer, Cham. https://doi.org/10.1007/978-3-319-60366-7_23

[7] Google Assistant, https://en.wikipedia.org/wiki/Google_Assistant, last accessed 2020/11/12.

[8] Kapko, M. (2018, February 7). Cortana explained: How to use Microsoft's virtual assistant for business. Computerworld, https://www.computerworld.com/article/3252218/cortana-explained-why-microsofts-virtual-assistant-is-wired-for-business.html, last accessed 2020/11/12.

[9] Siri, https://en.wikipedia.org/wiki/Siri, last accessed 2020/11/12.

[10] Cortana, https://en.wikipedia.org/wiki/Cortana, last accessed 2020/11/12.

[11] Alexa vs. Google Assistant - what's best for workplace apps?, https://www.adenin.com/blog/alexa-vs-google-assistant-whats-best-for-workplace-apps/, last accessed 2020/11/12.

[12] Amazon Echo, https://en.wikipedia.org/wiki/Amazon_Echo, last accessed 2020/11/12.

[13] Fallah, Navid, et al. "Indoor human navigation systems: A survey." Interacting with Computers 25.1 (2013): 21-33.

[14] Goshen-Meskin, D. R. O. R. A., and I. Y. Bar-Itzhack. "Observability analysis of piece-wise constant systems. II. Application to inertial navigation in-flight alignment (military applications)." IEEE Transactions on Aerospace and Electronic systems 28.4 (1992): 1068-1075.

[15] Alexa Skills Kit Glossary | Alexa Skills Kit. (n.d.)., https://developer.amazon.com/en-US/docs/alexa/ask-overviews/alexa-skills-kit-glossary.html, last accessed 2020/11/12.

[16] Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. Numerische Mathematik, 1(1), 269–271.

[17] Test Your Utterances as You Build Your Model | Alexa Skills Kit. (n.d.). Amazon., https://developer.amazon.com/en-US/docs/alexa/custom-skills/test-utterances-and-improve-your-interaction-model.html, last accessed 2020/11/12.

[18] Batch Test Your Natural Language Understanding (NLU) Model | Alexa Skills Kit. (n.d.). Amazon, https://developer.amazon.com/en-US/docs/alexa/custom-skills/batch-test-your-nlu-model.html, last accessed 2020/11/12.