How much could an ‘intelligent speaker’ such as Alexa or Cortana learn about a family’s lives and habits? How different would this be for a robot capable of moving around the home?

Introduction

Intelligent speakers are the realizations of conversational AI in homes. They utilize advanced deep learning functionalities like natural language understanding (NLU) and automatic speech recognition (ASR) to provide lifelike conversational interactions [1]. By design these devices are constantly listening and learning from multiple inputs to improve their user experience. This essay reviews how much Intelligent speakers learn about a family’s lives and habits. A study is made on the learning capabilities of Alexa which is by far the most popular Intelligent speaker till date [2]. A study is also made on Human Robot Interactions (HRI) to explore additional learning possibilities of an Intelligent speaker which is mobile. Various perceptions of robots are investigated and learning features are discussed.

Background

Alexa hardware mainly consists of multiple microphones that utilize noise cancellation and far field voice recognition so that it can pick up speech patterns from any direction [3]. Once Alexa is switched on its speech recognition software listens for a wake command from the user which is most commonly the word “Alexa”. It starts recording speech and the speech file is sent to the voice recognition service in the cloud. The voice recognition service interprets the speech and sends back the response to the intelligent speaker [4]. A study on the network traffic generated by Alexa revealed that it uses TLSv1.2 Encryption with certificate validation. According to the Washington Law Review the user speech files are stored in the cloud and this enables Amazon to associate recordings with the users Amazon account. While learning the users voice patterns to understand the user’s requests, Alexa also learns the user’s preferences with the use of ASR and NLU. ASR is a technology which converts spoken sound waves in to corresponding sequence of words. ASR is computationally intensive and utilizes layered mathematical functions modelled after biological neurons and statistical models to decide the right word in a situation of words having the same pronunciation [6]. Most mainstream ASR systems are designed as probabilistic Bayes classifiers which imply the need for vast volumes of training data to improve accuracy [7]. NLU on the other hand teaches computers to understand and interpret human speech based on grammar and speech’s context [6]. This involves converting the text generated by ASR into computer language and producing a response that humans can understand. NLU uses machine learning on past examples to disambiguate what the user meant from the words they spoke. Once again performance improvement of this technology solely depends on large corpus of historical transactions stored on the cloud.

Intelligent speaker Learning

Based on the arguments discussed about Alexa’s need to learn about the user let us now look in to how much Alexa would be able to learn from its users be it a single user or a family.

Voice commands: By default, Alexa is capable to answer questions or ask to follow up questions in the following areas [6]. These are also known as intents in the software stack:

- Books: Converses with users about books, their ratings and navigates through audio books
- Calendar: Answers users’ questions about calendars and upcoming events
- Cinema: showtimes
- General: For user requests that don’t fall in any of the listed categories
- Local search: Answers users queries about nearby businesses, opening hours and phone numbers

- Music: Plays music requested, creates playlists and music libraries.
- Video: Provides information about episodes of tv shows or any other video media
- Weather
- Built in intents: For controlling device functionality such as stopping, cancelling, or providing help.

General guidance from Amazon to Alexa software developers is: , “Alexa should remember context and past interactions, as well as knowing a customer's location and meaningful details in order to maintain familiarity and be more efficient in future exchanges” [10]. Conversation on a topic is modelled in software through ‘conversation flows’ which enable the developer to lead the flow of the conversation. This provides an opportunity to the software developer to channel the conversation in a way possible to learn more about the users preferences in any of the above areas.

Voice Authentication: Alexa is always listening for its wake-up word. In this process it may listen to other voices in the room such as TV programs and advertisements. DolphinAttack was an incident which proved Alexa listens to other voices including inaudible voices [9]. Since all speech files are processed in the cloud this feature could act a window into the user’s life for interested parties.

Company monitoring: There has been instances where Amazon employed manual processing of speech files. A Bloomberg investigation revealed that Amazon had given contracts to thousands of employees to process up to 1000 audio clips in 9h shifts. Another incident is where a customer decided to exercise his GDPR rights and requested his stored personal information from Amazon. Amazon mistakenly sent 1700 audio files and a transcribed document containing interactions of other users with the device These incidents shows that Amazon could learn about user preference and habits even manually and that user data in the Amazon cloud is up to Amazon’s disposal [9].

Misinterpretation: To improve learning Alexa also stores speech files of conversations it did not understand. These conversations can be found from the history setting of the device. A typical example of a failed conversation is the utterance of the word “Alexa” which was not meant for the device. The dollhouse incident is one such famous example [5]. This is another learning opportunity for the intelligent speaker to learn about the daily life of its user.

Skill personalization: Amazon provides their proprietary speech recognition technology to third party developers to enable them to build new applications for the Alexa device. These third-party applications are called skills. It allows companies to build customized voice applications tailored to the services they offer. Uber and Domino’s are examples of companies providing skills for Alexa. The skills are broadly divided into Custom skills for general use by wide range of applications, Smart home skills for IOT uses and Flash briefing skills for news summary. Note the wide range of user data Alexa is capable of handling. What is interesting about skills is that it allows a feature called skill personalization. In a home there maybe more than one user interacting with an Alexa device. Each user might have different needs and preferences. If the skill supports personalization it will be able to differentiate between users using their voice profile. For example, an exercise skill would be able to provide different exercise routines to different users interacting with the device. Similarly, with ride hailing, if the multiple users have linked their accounts with the device, the device can choose the account of the user with which it is interacting for hailing a ride. This feature enables the speaker to learn about the different members of the family and their preferences.

From the Amazon developer pages it is evident that Alexa recognizes users and provides this capability to third party developers for using it in skill development [8]. Amazon lays the responsibility of data privacy on the developer and the user in this case. This serves as a dual-use capability: for improving user experience as well as threat to data privacy.

Since Alexa is a cloud based storage system the amount of highly personal data it can collect is virtually unlimited and due to the limitations in cloud analysis by an external party it is hard to determine what kind of data gets stored and how it gets utilized. But based on the various incidents we can speculate the breadth and depth of the learning Alexa does.

Robot Learning

Next, an investigation is made into the learning possibilities of an intelligent speaker which can move around. Such devices are called social robots and the interaction with them is called Human Robot Interactions (HRI). The ways in which these devices would learn would be a superset on top of what was previously discussed. We start by looking into the additional perception methods available commonly in social robots and then we look in to features that can be supported by these perceptions. We identify what can be learned about a user or family using these features.

Perception methods in social robots can be classified in to visual-based, audio-based, tactile based and range sensor based [11].

Learning through visual perception: Through visual signals social robots will be able to detect faces, track humans, identify facial expressions and understand gestures. They will be able to identify colour, shape, and texture of objects. In case 3d visual signals are available from Kinect or stereoscopic cameras, they can be used to create depth images which provide distance learning capability to the robot. The robot will be able to locate a particular object and learn about its surroundings.

Learning though audio perception. This has been discussed earlier in the context of the Alexa intelligent speaker. For a social robot it will have the advantage that the additional perception methods will act as learning cues for the audio perception and performance of this learning method will be improved significantly. We discuss about this in multimodal perception later.

Learning though tactile based perception: Through tactile perception social robots will be able to learn the humans emotional state to a certain extent. For example, a light pat would mean the user is relaxed and heavy scratch would mean the user is in an angry state. In addition, more information can be gained from the touched objects depending on the type of sensor that was used. Temperature of the user is one thing that can be learned through this method.

Learning through range-based perception: Social robots could use Laser range finders to track their users and identify habits. 3D range finders will enable robots to explore the whole a 3D scene around it.

Learning through multimodal perception. Signals from the above perception methods can be fused to address shortcomings of relying only a single perception method. Multimodal fusion is also applied to develop the attention system of the robot. It is also used in audio visual based emotion recognition.

The perception methods discussed enable social robots to respond to human initiated inputs. Pepper is an example of an interactive social robot which recognizes faces and human emotions. Robots can improve their social interaction logic by learning human speech and motion behaviour through imitation. Another approach to improve interaction logic would be for robots to learn proactive behaviour from human interaction data. Robots will be able to identify opportunities of proactive action to be generated based on the user interaction history. For example, a robot was trained to behave like a proactive shopkeeper in a camera shop in the following proposal [12]. The result of the experiment was that the robot was not only able to answer the question raised by the customer, but also proactively assist the customer in introducing new features or a new camera. Such proactive behaviour enables a social robot to both learn as well as influence the lives of the family in a home.

Conclusion

As discussed so far, intelligent speakers learn about a family’s lives and habits in several ways and this improves with the advancements in algorithms for speech recognition and natural language understanding. The learning is augmented with the availability of vast amounts of data with more customers purchasing intelligent speakers. As of 2019 there are 133 million smart speakers in use and 100,000 Alexa skills available worldwide [6]. Alexa was able to understand 99% questions it was asked and provide correct answers to 79.8% questions it was asked. A study on the usage patterns of Alexa indicated that the second most common location of the device was the bedroom [3]. This implies that despite all privacy concerns users have approved of the device to a great extent. So, based on the understading that the device learns about its family to provide better experience, a few recommendations are made which will help to secure the learning process of the device and improve user privacy and confidence in the device [9].

Enable device to work without cloud: Most of the security concerns can be directed towards sending of every user input to the cloud for processing. This could be avoided if the learning capabilities could be implemented in the device hardware itself. This would require significant upgrade of the device hardware and a proportionate increase in device cost. But advances in speech recognition algorithms gives us hope of cheaper hardware requirements in the future.

Incognito mode: It was found that some users approved of intelligent speakers using stored information to improve performance. But in certain instances, they wished the speaker could just skip listening. Web browsers provide this feature through incognito mode or in private mode. A similar feature in intelligent speakers would enable users to be more comfortable using the device.

User education and improved controls: Users should get to know the big picture of how the device works and how their data is processed. There is no information available on the internet which can be checked though a normal search. Users should be made aware that manual listening of their conversations could happen as cited earlier. Users should be provided controls in the device to accept or reject this practice and the controls should be easily reached.

Muting the device through voice: Even though there is a physical mute button on the device, it is an inconvenience to walk to the device to mute it every time. In case of disabled users this is more of a limitation of the device. Giving the users convenient and absolute control over when the device listens or not will help to build trust in the device.

In this essay we discussed about the how much intelligent speakers could learn about a family and envisioned how much more could be learned if the speaker were mobile. We investigated the different perceptions of available social robots and understood their learning capabilities and future. Finally, we proposed a few recommendations towards securing the use of intelligent speakers in a home environment.

References:

[1] V. Këpuska and G. Bohouta, "Next-generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home)," 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, 2018, pp. 99-103, doi: 10.1109/CCWC.2018.8301638.

[2] Sciuto, Alex, Saini, Arnita, Forlizzi, Jodi, and Hong, Jason. ""Hey Alexa, What's Up?"." (2018): 857-68. Web.

[3] Lopatovska, Irene, Rink, Katrina, Knight, Ian, Raines, Kieran, Cosenza, Kevin, Williams, Harriet, Sorsche, Perachya, Hirsch, David, Li, Qi, and Martinez, Adrianna. "Talk to Me: Exploring User Interactions with the Amazon Alexa." Journal of Librarianship and Information Science 51.4 (2019): 984-97. Web.

[4] Zhang, Xiaolu., and Kim-Kwang Raymond. Choo. Digital Forensic Education : An Experiential Learning Approach. 1st Ed. 2020. ed. Cham: Springer International : Imprint: Springer, 2020. Studies in Big Data, 61. Web.

[5] Liptak, Andrew. “Amazon’s Alexa Started Ordering People Dollhouses after Hearing Its Name on TV.” The Verge (January 7, 2017). https://www.theverge.com/2017/1/7/ 14200210/amazon-alexa-tech-news-anchor-order-dollhouse.

[6] Shih, Win. "Voice Revolution." Library Technology Reports 56.4 (2020): 5. Web.

[7] Virtanen, Tuomas., Rita. Singh, and Bhiksha. Raj. Techniques for Noise Robustness in Automatic Speech Recognition. Chichester, West Sussex, U.K. ;: Wiley, 2008. Web.

[8] https://developer.amazon.com/en-US/docs/alexa/custom-skills/add-personalization-to-your-skill.html (Accessed: 02 Nov 2020)

[9] Moallem, Abbas. HCI for Cybersecurity, Privacy and Trust. Cham: Springer International AG, 2020. Web.

[10] West, Emily. "Amazon: Surveillance as a Service." Surveillance & Society 17.1/2 (2019): 27-33. Web.

[11] Yan, Haibin, Ang, Marcelo H, and Poo, Aun Neow. "A Survey on Perception Methods for Human–Robot Interaction in Social Robots." International Journal of Social Robotics 6.1 (2013): 85-119. Web.

[12] Liu, Phoebe, Glas, Dylan F, Kanda, Takayuki, and Ishiguro, Hiroshi. "Learning Proactive Behavior for Interactive Social Robots." Autonomous Robots 42.5 (2017): 1067-085. Web.