A Survey of the Rise of the Virtual Assistant

Jayro Alvarez
Department of Computer Science
California State University, Fullerton
Fullerton, California 92831, USA
jayroalvarez@csu.fullerton.edu

Zulema Perez
Department of Computer Science
California State University, Fullerton
Fullerton, California 92831, USA
rancid80s@csu.fullerton.edu

Patrick Myers
Department of Computer Science
California State University, Fullerton
Fullerton, California 92831, USA
pmyers@csu.fullerton.edu

Adam Shirley
Department of Computer Science
California State University, Fullerton
Fullerton, California 92831, USA
bsa919adam@csu.fullerton.edu

***Abstract*—This survey reviews the current increase in popularity of virtual assistants - intelligent agents that can transcribe spoken natural language and provide a synthesized response and, or action. In this document we provide a historical overview of chatbots to virtual assistants, we review the standard dialogue system architecture of a virtual assistant. We describe the pipeline framework and components of the dialogue system, briefly touching on non-goal-oriented dialogue systems called chatbots and discuss goal-oriented dialogue systems which can also be called virtual assistants. We also review the current applications of virtual assistants, the possible future applications of virtual assistants, and potential security issues and privacy issues of virtual assistants and general use cases of virtual assistants.**

I. Introduction

According to an article on ScienceDirect [1], Alan Turing introduced the idea of a "thinking machine" in 1950 in a paper called "Computing Machinery and Intelligence." In his paper he asks, "Can a machine think?" And he describes the question as absurd as it would require the definitions of "machine" and "think" be perfectly defined, despite their ambiguous nature. He then proposes an idea that a 'game' will be played between a man, a woman, and an interrogator of any gender. The interrogator will prompt a teleprinter to ask the question to the individuals, without being able to see either outside a label and who will whisper said answer into the ear of the teleprinter (to prevent tone of voice from giving them away), and the teleprinter will offer a typed response to the interrogator. Neither the man or woman will be able to see the interrogator. One individual will attempt to fool the interrogator, and the other tries to convince him of who they are.

The game will progress until the interrogator is either fooled or can understand which is which. He then says, let's replace one of them with a machine. If hypothetically speaking the machine can play this game as well as a human, it is an accurate replacement.

As Ishida contemplates, to perfectly define whether or not something can "think" we also have to give an abstract understanding of questions such as self-awareness and free will, both of which we cannot fully understand ourselves.

The information the article has described is relevant as the chatbot attempts to do just this. An algorithm in a machine can help describe "free will" using nodes that effectively represent conflicting thoughts in a machine. Does a chatbot think? To an extent, one could argue it does. If you were to hypothetically avoid introducing one as a chatbot and it's able to chat with the human regularly, it could pass such a test.

A more recent development has been a comparatively less intelligent, more purpose-specific development known as virtual assistants. According to a Cornell research article by Radzwell [2], the first chatbot that could even be expressed as a chatbot was ELIZA in 1966. We establish here the major difference between a chatbot and a virtual assistant. A virtual assistant will appear more intelligent until you ask it about something it is not meant to know about. As an example, if you were to ask Siri "what's the weather today?" She'd respond with a detailed hour-hour estimation of the weather for about 12 hours by accessing your location and referencing it on a website "weather.com." If you were to ask her about her dreams, she might simply link you to a Wikipedia article about a book called dreams. If you ask her "why" she doesn't retain information about the previous response she may have given. A generic response such as "I don't know why" will be given. With a chatbot, asking them about the weather will result in responses, usually along the lines of "I don't know," or "rain I guess?" Asking about their dreams will likely result in them giving a more responsive answer and asking them something like "why," depending on the chatbot may result in the chatbot attempting to elaborate its previous answer. A virtual assistant is purpose-specific, a chatbot is a general conversational agent.

Eliza was originally developed with the intent to create a virtual Rogerian psychotherapist inspired by Turing's paper on "Thinking Machine." According to Deshpand research [3] these bots were not considered very intelligent, instead of using a series of predefined interaction based on specific inputs. It details the natural evolution these bots went through.

Parry (1972) came next called ”Eliza with an attitude” meant to stimulate paranoia. It passed the Turing test 52 percent of the time. Jabberwacky was more focused on entertainment but played an important role by opening the gate to voiced AI. In 1995 Alice used a much more advanced heuristic method to understand input but had limited output based on predefined responses.

Through continued interaction, primitive designs such as ELIZA were never expected to pass the Turing test, but only attempt to give minimal effort responses to progress natural learning. Cleverbot is a chatbot that was developed in 1986 and put online in 1997. Since then we've had many publicly available chatbot AIs such as Microsoft's Tay (was taken down when it learned racism), Microsoft's Zo, Replika, Sophie count as such. From this, virtual assistants such as google assistant using a technology called duplex, Siri, Alexa have been developed.

While both chatbots and virtual assistants act as conversational software agents utilizing natural language input, virtual assistants themselves are useful for very specific commands, but generally, aren't useful for a complex discussion or conversation. The article this information has come from is relevant as it describes the history of our AI development using a slowly progressing network of expanding agents leading to modern issues. The same article notes that many modern influences of our current life can be described with these issues and concerns. As an example the article describes, ”in the 2016 election, semi-autonomous bots made a large volume of posts on places like facebook and twitter. Cleverbot as a chatbot works by interpreting the repeated patterns in a conversation between two individuals. According to a research article by IJESC [4], Virtual assistants are something closer to search engines that look for specific voice commands and associate them with functions specific to the device. They are becoming progressively more intelligent due to recent advances, ”While their principle work is to react to directions, in doing as such, they additionally learn. The more a man interacts with voice-actuated gadgets, the more patterns and examples the framework recognizes depending on the data it gets. ” The article in question is relevant as it describes the current and ongoing expanding way we define virtual assistants. The four major ambassadors of virtual assistants in modern computing are now Apple's Siri (2011), Microsoft's Cortana (2014), Amazon's Alexa (2014) and Google Assistant (2016) [5].

II. Dialogue System Architecture

Several areas of speech, natural language processing, and natural language understanding have had substantial breakthroughs contributed by the development of big data and deep learning techniques [6] over the last couple of decades. In the area of dialogue systems, deep learning techniques have maximized usage of an enormous amount of data to gain an understanding of important feature learning, representation learning techniques, and neural response generation procedures while requiring the least amount of engineering [7]. Dialogue systems are categorized into two groups; non-goal oriented dialogue systems and goal oriented dialogue systems, also known as non-task oriented dialogue systems and task-oriented dialogue systems. Chatbots are non-goal-oriented dialogue systems, and virtual assistants are goal-oriented dialogue systems. In this document, we will focus on virtual assistants and the goal-oriented dialogue system.

A. Non-Goal-Oriented Systems

Non-goal-oriented dialogue systems have been around since the mid-1960s. As mentioned earlier, Eliza, a popular program of the time, was a non-goal-oriented dialogue system based on basic textual parsing rules. Non-goal oriented dialogue systems are designed to interact with users and provide mental stimulation and support [8]. In this survey, we will focus on goal-oriented dialogue systems.

B. Goal-Oriented Systems

Goal-oriented dialogue system work mostly based on deterministic design and engineering guidelines combined with speech recognition [9]. Machine learning techniques were later used to categorize the need of the user and to close the gap between text and spoken words. In the mid-1990s, based on Markov's decision-making process, research in goal-driven dialogue systems grew in popularity and researchers started to develop dialogue as a sequential choice making problem. Goal-oriented dialogue systems are designed to assist users to complete certain tasks through understanding the user's requests and providing the requested information [10].

III. Pipeline Framework

The standard pipeline framework of a text-based goal-oriented dialogue system consists of four main components natural language understanding, dialogue state tracking, dialogue response-action selection, and natural language generating. For a speech-based goal-oriented dialogue system, two more components are incorporated into the pipeline framework, automatic speech recognition, and text-to-speech synthesizer [11].

- Automatic Speech Recognition - Automatic speech recognition [12] converts the user's spoken word(s) into text the computer can understand to be identified and processed by the natural language understanding component.

- Natural Language Understanding - After the automatic speech recognition component passes the converted text, the natural language understanding [13] component parses the text into organized blocks of already established categories. The categories are based on the users objective and multiple scenarios.

- Dialogue State Tracking - The dialogue state tracker [14] attempts to guess the users objective at every turn of the dialogue. A dialogue state indicates the representation

of the dialogue period. The dialogue state tracker is commonly called a slot filler or systematic frame.

- Dialogue Response-Action Selection - Based on the dialogue state representation of the dialogue state tracker, the dialogue response action selection [15] component must select the correct dialogue system response or action. The dialogue response action selection component also learns the next action or response from the current dialogue state.

- Natural Language Generating - The natural language generating [16] component converts and maps the selected dialogue system response or action into the natural language the user understands.

- Text-to-Speech Synthesizer - Text-to-speech [17] synthesizes the spoken words produced by the dialogue system into natural language then reads it to the user.

This section will further discuss the four main components of the pipeline framework.

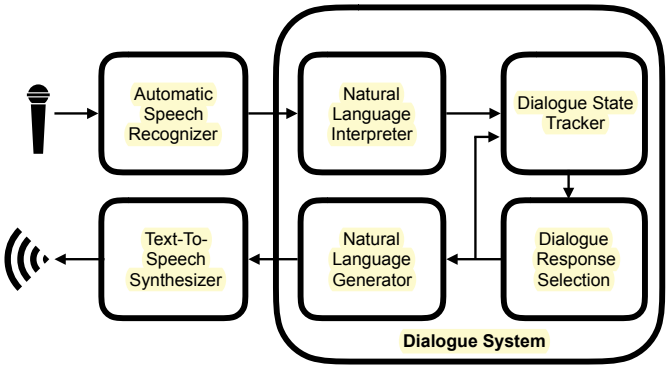

Fig. 1. Dialogue System Diagram - [15] The standard architecture of a speech-based dialogue system pipeline framework for a goal-oriented dialogue system consists of four main components natural language understanding, dialogue state tracking, dialogue response-action selection, and natural language generating and two additional components automatic speech recognition, and text-to-speech synthesizer.

A. Natural Language Understanding

Given some spoken words, the natural language understanding maps the spoken words into logical slots. The slots are predefined depending on different scenarios [18]. Usually, there are two types of representations. One is the spoken word category, which deals with the user's purpose and the spoken word category. The second is the world-level information extraction like recognition of named entities and filling slots. The purpose detection is performed to detect the intent of a user. It categorizes the spoken words into one of the predefined intents. Deep learning techniques have been used for purpose detection. Filling the slots is a challenging problem for natural language understanding [19]. Unlike purpose detection, filling slots is usually defined as a series labeling problem, where words in a sentence are assigned semantic labels. With the input being a sentence of a series of words, and the output is a series of slots, one for each word. Furthermore, the logical representation created by the natural language is further processed by the dialogue management component. The average dialogue management component has two stages, a dialogue state tracker and dialogue response action selection [14].

B. Dialogue State Tracking

Tracking dialogue states is the main component that makes a dialogue system strong. It estimates the user's purpose at every word of the dialogue. The traditional structure is usually called slot filling or a logical frame [20]. But these rule-based systems have frequent errors because the most common result is not always the result that is desired.

C. Dialogue Response Action Selection

Determined by the state representation from the state tracking component, the dialogue response-action selection is to create the next available system action[21]. Either supervised learning or reinforcement learning can be used to optimize dialogue response-action selection. Commonly a rule-based agent is used to warm and start the system.

D. Natural Language Generating

The natural language generating component converts an abstract dialogue action into a natural language spoken word output. A good natural language generator commonly relies on , adequacy, fluency, readability, and variation. Traditional approaches to natural language understanding usually perform sentence planning [22]. It maps logical input symbols into the form representing the spoken words like template structures and converts the structures into final responses through the surface realization.

IV. End-To-End Systems

Goal-oriented dialogue systems have limitations. The first limitation is that the typical goal-oriented dialogue system uses a pipeline framework structure that connects the core dialogue system's components, which makes it difficult to identify error sources and to optimize system targets. The second limitation is that the input of one component is dependent on the output of another component [23] so when modifying one component and implementing it in a new area or introducing new data, all the other components will also need to be modified or introduced to the new data to ensure global optimization [24]. Blocks of categories called slots and feature might change which requires a great amount of effort. With the advances in end-to-end neural generative models in current years, many attempts have been made to construct an end-to-end trainable framework for goal-oriented dialogue systems. Instead of the classic pipeline framework, the end-to-end model uses a single module and interacts with structured external databases. [25] and [18] introduced a network-based trainable goal-oriented dialogue system, which converted dialogue system learning as a problem of learning a mapping or slot filling from dialogue

TABLE I
AREAS OF RESEARCH IN GOAL-ORIENTED DIALOGUE SYSTEMS

Area	*Research focus*	*References*
Neural dialogue systems	Deals with drawbacks of modularized goal-oriented dialogue systems with a neural dialogue system that can interact with a structured database. Reinforcement learning based dialogue manager that offers robust capabilities for handling issues caused by other components of the dialogue system.	Li, 2017
Adding memory to goal-oriented dialogue systems	Deals with the role of memory in goal-oriented dialogue systems. Based on frames and introduces frame tracking, which extends the state tracking component to where several states are tracked at the same time.	Asri, 2017
Ethical challenges in dialogue systems	Deals with the implicit bias in dialogue systems, possible sources of privacy violations, safety concerns, special treatment of reinforcement learning systems.	Henderson, 2018
Deep learning for dialogue systems	Deals with the advancement in deep learning technologies and the rise of applications of neural models to dialogue systems. Describes recent research for building dialogue systems and provides an overview of dialogue system challenges and improvements.	Chen, 2017
End-to-end optimization of goal-driven and visually grounded dialogue systems	Deals with encoder-decoder architectures for sequence-to-sequence learning. Introduces reinforcement learning method to optimize visually grounded goal-oriented dialogue systems based on the policy gradient algorithm.	Strub, 2017
Dialogue learning with human teaching and feedback in end-to-end trainable goal-oriented dialogue systems	Deals with a hybrid learning method for training goal-oriented dialogue systems through online user interactions. Addresses efficiency issues with a hybrid imitation and reinforcement learning method. This is accomplished with the design of a neural network based goal-oriented dialogue agent that can be optimized end-to-end.	Lui, 2018
Key-value retrieval network for goal-oriented dialogue	Deals with neural goal-oriented dialogue systems that struggle to smoothly connect with a knowledge base. Addresses the problem with a new neural dialogue agent that effectively sustain grounded, multi-domain discourse through a simple key-value retrieval component.	Eric, 2017
Adversarial learning for neural dialogue generation	This article draws upon the Turing test, using adversarial training for open-domain dialogue generation. Along with adversarial training a model for adversarial evaluation that uses successes in fooling an adversary as a dialogue metric and also while avoiding potential pitfalls.	Li, 2017
Network-based end-to-end trainable goal-oriented dialogue system	Deals with current development in goal oriented dialogue systems requiring to create multiple components. Introduction of neural network-based text-in, text-out end-to-end trainable goal-oriented dialogue system and a new way of collecting dialogue data based on a simple pipe-lined Wizard-of-Oz framework.	Wen, 2017
Frame tracking model for memory-enhanced dialogue systems	Deals with resources and tasks of state tracking in dialogue systems. Frame tracking tasks require multiple frames, one for each user goal set during the dialogue. Proposes a model that takes as input, a list a frames created during the dialogue, and the dialogue acts, slot types, and slot values. The model out preforms a previous proposed rule-based baseline.	Schulz, 2017

histories to system responses and applied an encoder-decoder model to train the whole system.

V. HYBRID FRAMEWORKS

Combining neural generative and retrieval based frameworks can have important effects on goal-oriented dialogue system performance. [26] and [27] tried to combine both methods. Retrieval based methods commonly provide accurate but straight-forward answers. Generation-based systems provide expressive but unintelligible responses. In a combined model the advantages of a retrieval-based and generation-based models showed impressive performance.

VI. APPLICATIONS OF VIRTUAL ASSISTANT

In this section we discuss the current research and applications of virtual assistant in areas of industry and academia.

A. Impact of Conversational Style of Virtual Assistants

Different benefits arise when comparing social- versus task-oriented interactions between humans and virtual assistants. According to another article on ScienceDirect [28], the change of conversational style had significant interaction effects on a group of experimental test subjects. The differences came from two uniquely chosen divisions of test subjects; the test subject groups were grouped by Internet competency, low versus high. From a laboratory experiment, results revealed

TABLE II
AREAS OF RESEARCH AND APPLICATIONS IN VIRTUAL ASSISTANT

Area	*Research focus*	*References*
Technology behind virtual assistants	Deals with the natural language understanding, speech recognition, machine learning and structured data. Provides an overview of virtual assistants, and describes the system architecture, key components, and technology behind the virtual assistant and discusses the future of human-computer interaction.	Sarikaya, 2017
Skill discovery in virtual assistants	Deals with skill discovery in virtual assistants. Discusses how voice is the primary means of engagement, and how voice-activated virtual assistants are growing in popularity. Discusses about headless devices like Amazon Echo. Discusses that despite the value of of the virtual assistant, discovering all of the capabilities is still a challenge.	White, 2018
Managing uncertainty in time expressions for virtual assistants	Deals with the issue when people speak to plan out and schedule things they often express uncertainty about time and use vague expressions. The modern virtual assistant often lacks the system support to capture the intent behind this vague time expressions, which can result in erroneous scheduling. The article ends with suggestions on design for future virtual assistants.	Rong, 2017
Next generation of virtual assistants	Deals with the overview of current virtual assistants on the market and how the next generation of technology based on the multi-modal dialogue system which can process two or more combined user input modes. Wants to create a movement to design the next generation of virtual assistant models using multi-modal system designs.	Kepuska 2018
Natural language processing for industrial applications	Deals with the challenges and limitations of industrial virtual assistant applications, with focus on the ”human in the loop” aspect. Cooperation between human and machines as a mutual interest.	Quarteroni 2018
Introduction to virtual assistants	Deals with virtual assistants and provides an overview of virtual assistants. Discusses what virtual assistants are, what they can do, security and privacy of virtual assistants and potential future uses.	Hoy, 2018
Continuous authentication for virtual assistants	Deals with continuous authentication for virtual assistants. Discusses the difficulty in securing virtual assistants and the dangers of having lack of security on virtual assistants. Discusses research that could solve the security issues that coincide with not having continuous authentication for a virtual assistant.	Feng, 2017
Almond: The architecture of an open, programmable virtual assistant	Deals with the architecture of Almond, a programmable virtual assistant for online services and the Internet of Things.The research addresses four challenge sin virtual assistant technology with a working prototype. Almond is the first virtual assistant that lets users specify trigger-action tasks in natural language.	Campagna, 2017
How an artificially intelligent virtual assistant helps students navigate the road to college	Deals with deep reinforcement learning using convolutional neural networks which is the technology behind autonomous automobiles and implementing it into a virtual assistant application that can help students navigate matriculation into college.	Page, 2017
Virtual assistants and self-driving cars	Deals with the personification of the car intelligence incorporating an algorithmic brain, a synthesized human voice and sensor-based senses. Poses the question, should virtual assistants assist or replace humans whenever necessary.	Lugano, 2017

that users' Internet competency and the virtual assistant's conversational style had significant effects on social, functional, and behavioral tendencies.

The purpose of choosing these specific parameters was to analyze the differences between groups pertaining to older and younger demographics. The results of this experiment gave significant results pertaining to the sample size of older participants. Specifically, this experiment found that it may not be smart to design a virtual assistant with a 'one-size-fits-all' strategy due to the reactions of the less competent, older participants. It was found that these older participants were more hesitant to trust a computer since there was a direct correlation between low Internet competency and distrust in computers [28]. This is why a virtual assistant designed with a more inviting and happier conversational style is preferred for older users, to generate a form of trust.

B. Virtual Assistants Built on Basic Emotion Theory

A possible fundamental building block in the creation of virtual assistants may revolve around the idea of basic emotions theory. The paper on the topic, by Zhiliang Wang, goes on to discuss how basic emotions theory points out that compound emotion consists of eight major basic emotions and how different combinations of these major basic emotions can be interpreted as a reflection of human will [29]. This could

be important to developing a virtual assistant with similar decision-making processes of a human.

By manipulating parameters of basic emotions in our system, it may be possible to create a virtual assistant which obeys human emotion rules. This may be the future of some virtual assistants, some that may have their own emotion.

C. Almond: An Open-Source Programmable Virtual Assistant

Almond is an open-source programmable virtual assistant. An understanding of Almond is crucial to the topic of virtual assistants because it gives insight into the basic architecture of one of many virtual assistants. There are four main areas of interest regarding Almond: generality, interoperability, privacy, and usability. Generality is achieved through Thingpedia, and a crowdsourced public knowledge base of open APIs and their natural language interfaces. Interoperability is achieved through ThingTalk, a high-level domain-specific language that connects multiple devices or services via open APIs. To achieve proper privacy, user data is managed by the open-source ThingSystem which can be run on personal phones or home servers. Usability is achieved through a natural language interface which allows users to specify action tasks in a natural language [30].

Taking a more in-depth look at Almond's main architectural components gives us a better understanding of how developers can add new functionality to the system and a better general view of how the components interact.

As previously stated, generality is achieved through Thingpedia. More specifically, Thingpedia is an encyclopedia for the Internet of Things or IoT. Every device that is compatible with Almond can be found in Thingpedia, along with the natural language interface and necessary specifications corresponding to a device's application programming interface or API [30]. Thingpedia is an essential component which allows developers to add to the Almond system.

Also as previously stated, interoperability, or the ability of computer systems or software to exchange and make use of information [31], is achieved through ThingTalk. ThingTalk is a high-level language created to connect IoT. Through ThingTalk, it is possible to connect the APIs of devices without exposing the specifics going on behind the scene such as the details of configuration [30]. ThingTalk is essential in taking the virtual assistant's commands, from the natural language of the user, and knowing what set of rules to apply to them based on the existing knowledge of devices located on Thingpedia.

The final important component of Almond to go over is ThingSystem. As previously stated, ThingSystem allows Almond to run under proper privacy. While Thingpedia is responsible for containing all universal information about Almond-compatible devices, each user has their ThingSystem which stores information about their connected devices. ThingSystem is very light-weight and can run locally on a user's phone or up in the cloud. The main roles of ThingSystem are to provide device management and configuration to the user and also execute the ThingTalk commands that are produced by the virtual assistant after having interacted with the user [30].

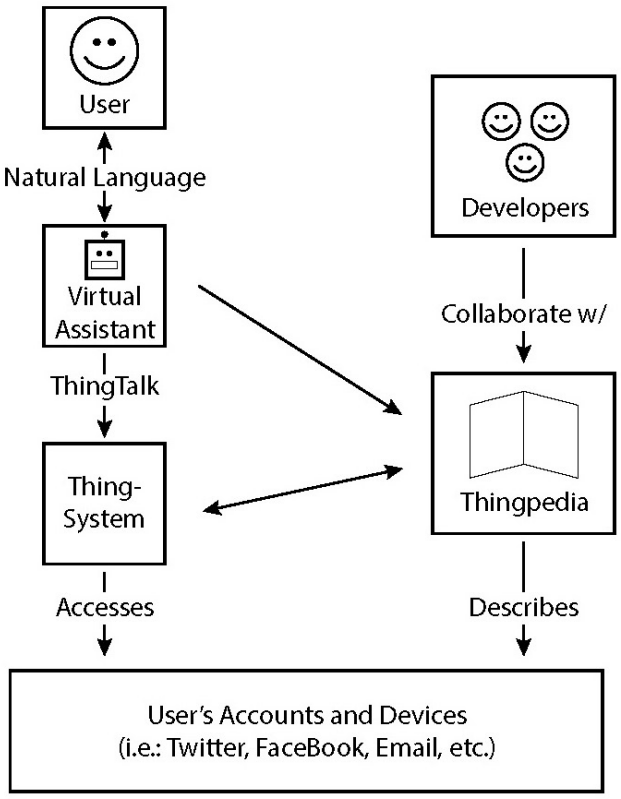

Fig. 2. Almond Architecture - [30] The architecture for Almond is built on multiple components interacting with each other in order to communicate and complete tasks with a user's accounts and devices. The specific technologies are explained with greater depth within the paper. The figure shows a general sense of the interconnectedness of the Almond system.

Almond unlocks a wide variety of access to many devices from just one interface. In other words, this means that through Almond, a user can complete a variety of tasks, usually requiring going into different applications, by just giving commands to Almond. The commands accepted by Almond can be categorized into two classes of operations: primitive and compound [30].

Primitive commands are some of the most basic commands Almond accepts; these commands include direct actions or queries to the device. An action could be a command to send a text to someone, and a query could be a command to ask for the name of a missed phone call. Primitive commands also include standing queries, known as monitors, and filtered monitors. An example of a monitor would be a command to have Almond let a user know when they have received a call. A filtered monitor is a monitor with added user-defined triggers. An example of such can be seen with adding the filter of a specific contact name to the previous example of Almond notifying a user of a received call.

Compound commands are more complex commands which may involve two or more functions compacted into a single unit for Almond to process. Typically, compound commands

Class	Type	Example
Primitive	Action	Send a text to Mary
	Query	Get my latest missed call
	Monitor	Let me know when I get a call
	Filtered Monitor/Query	Let me know when I get a call **from Tom**
Compound	Trigger + Query	Everyday show me the weather at 7am
	Trigger + Action	Everyday play music at 6:30am
	Query + Action	Get my newest picture and send it to Mary
	Trigger + Action + Query	Everyday send current weather to mom at 8am

Fig. 3. Almond Command Types Chart - [30] Almond categorizes commands into primitive and compound. The chart summarizes and gives example of the types of commands that can be set by users.

may include a multitude of different combinations of primitive commands. One example can be seen with the combination of a monitor, query, and action in a command to Almond demonstrated in the following: **monitor** the time for 9 P.M., **query** for any rainfall tomorrow, and alert (**action**) the user whether they will need an umbrella.

VII. General to Specialized Use Cases

In this section we discuss how different subgroups use virtual assistant applications to meet their needs. In this section, we will discuss how different people interact and use voice assistants, their expectations of these interactions entail and how these contrast from different user groups. First, we look at a study conducted in Bangalore.

VIII. General Use Cases

A. Why and when people use voice assistant conducted in India

This article is about a survey conducted in Bangalore on when and how people in a low economic group and a middle economic group use voice assistants and voice search functions. The study found that for the people in the low economic group they used the voice search and voice assistant mainly if they were only semi-literate and they used it only for searches and no other purposes. The middle economic group, on the other hand, used less of the voice search in general due to being more literate while at the same time probably due to being more tech savvy they would use the assistant features to set alarms and reminders. The study also showed that form both groups those trying to learn English would often type to practice and have Google correct any spelling mistakes so they could more effectively learn to spell [32].

This article demonstrates how the voice assistant technologies can be used to help people who are not literate to still be able to easier access the internet and give a basis for how this concept to be expanded to people with other problems from disabilities to seniors who have trouble typing.

B. Voice Interfaces in Everyday Life

This article is about a study that was conducted about how people interact with an Amazon Echo in a natural Conversational Setting. The article describes how a family who participated in the study interacted with their echo in different ways. They make a number of different observations about the dynamic the family uses to interact with the echo from things like how the people would try to help formulate requests or questions that the echo failed to recognize and answer properly in turn based fashion where one person tries followed by another [33]. In the article, the main way the family interacted with the echo was through various skills in the form of different kinds of quizzes. In these interactions, the authors note that the echo is not treated as a member of the conversation itself and instead is more something the conversation is based around or used to supplement information for the conversation rather than the people in the family talking specifically to the echo. This contrasts to the way that subjects that took part in the cognitive impairment article which will be discussed later because in that article the participants explicitly stated that one of the things they would like to be able to use the voice assistants for would be as a conversational partner [34]. Another hope that the subjects of this article mentioned were that the virtual assistant would assist them in their interactions with other people. They proposed this be done by making it so that the virtual assistant would be able to help translate their speech to be more understandable by other people so that they could more easily interact with others.

C. Compare and Contrast

These two studies previously mentioned can give us a good look at the diversity of the people who use virtual assistants. On one hand their is the users in Bangalore who from the context of the article took a more practical use to the virtual assistants with them focusing on the features that made the users in each the middle economic group and the low economic group both used features that helped to streamline their phone usage as was mentioned above. On the other hand, there is the Canadian family with the echo who seemed to use the device mostly for recreational purposes such as playing games. This contrast is likely made to several factors such as the environment the two different groups live in as well as of course the difference in the virtual assistants that they were documented using.

IX. Specialized Use Cases

A. Specialization of Virtual Assistants

This brings us to a point about the specialization of the different virtual assistants. The two examples we have Alexa being utilized on an Amazon echo and Google Assistant on a smart phone. The two devices themselves while having some overlapping uses are very different in both form and function. These two devices both can be used for simple things like setting the alarm or searching the internet based on a question spoken aloud. Where they differ is where it is unlikely anyone would play a quiz game utilizing their phone with Google

Assistant while with Alexa on an echo a prime example was given in Porcheron's study. This opens up many possibilities with the possible specialization of different virtual assistants possibly on different devices for vastly different purposes.

B. Helping older adults with intelligent health care assistant

This article explored a more narrow field of possible specialization for personal assistant being health care assistance specifically for older adults. Their research was mainly a simple survey to see what interest older adults would have in this area [35]. The paper's premise was that the user would feed the system there personal health care information and then provide would advise based on this information. The paper gave three options for which kind of advice the user would want. The three being a general heal information search such as what you would do currently on the internet, a personalized health information search which would take all of the users entered data past and present into account to provide a recommendation, and lastly a fully integrated health regimen which would be the system giving the user tips on what to do on a daily basis to improve their health.

This article is an example of how voice assistants could be used for many different specialized purposes like giving medical advice. There are many possibilities for such specializations that could help people improve the way they live from a voice assistant doing personalized workout routines to something to help a student schedule their study time. This is a definite area that developers should look to further expand so that these virtual assistants can help to improve the quality of people's lives. Another more specialized case mentioned in an article in the next section is about people with cognitive impairments in which they mention the idea of the virtual assistant on their phone being able to give commands to vending machines making their lives easier [34]. This just goes to further emphasize the possibility of giving virtual assistants more specialized capabilities that utilize existing technology because as the researchers for those articles mention there already existing technology to pay vending machines with utilizing a smartphone so it would just require new interfaces to be designed to make this possible rather than new hardware.

C. Use of Voice Activated Interfaces by people with Intellectual Disability

This article studies the possible applications of voice searches and assistance in helping people with Intellectual disabilities. The definition for an intellectual disability which the article referenced was "a disability characterized by significant limitations in both intellectual functioning and in adaptive behavior, which covers many everyday social and practical skills. This disability originates before the age of 18" [36]. The article goes on to talk about the results of the study which were overall very positive with 50% of participants being able to complete the three different tasks that they used for this research, and 55% were able to complete two of the three tasks [37]. After the researches completion, 72% preferred using voice search options rather than typing.

This study shows us many possibilities for the use of voice assistants to help those with intellectual disabilities or even just people that have a harder time with technology. By making it easier to use and reducing the buttons that need to be pressed it is possible to help people that struggle with technology due to the complexity or confusing button placements by simplifying it with the use of searches done by speech. This study demonstrates this concept and mentions it in depth with talking about how the participants liked that the device gave audio feedback and how they would be confused when it did not respond verbally.

This study also shows where possible usability improvements could be made such as mentioned previously making sure that the voice assistant responds verbally in all cases whether it times out, fails to understand what the user is saying, or when it completes the search or action. The article also mentions having the ability to set a usability feature that would allow for longer times for inputs as well as recognizes slower speech patterns which according to Balasuriya is a common trait for people with Intellectual Disabilities as well as the elderly [37]. Some other minor suggestions were to make the buttons stand out more, reduce the number of distractions on the page, and better able to recognize what people with speech impediments are trying to say.

D. Use of Voice Assistants by those that are cognitively impaired

Another Article similar in nature to the previously mentioned one is an article which explored the options for the use and the expectations that people who are cognitively impaired have for virtual assistants. According to the CDC "Cognitive impairment is when a person has trouble remembering, learning new things, concentrating, or making decisions that affect their everyday life "[38]. So basically a person with a Cognitive disability is someone who may or may not have trouble understanding information but may depend on the degree of severity be able to apply and use this information while someone with an intellectual disability will have trouble applying the information but not necessarily acquiring it. This article talks about what these people who have cognitive disabilities would like from a virtual assistant as well as of course any issues they may have with the virtual assistants. One of the main points in the article was the necessity for robust speech recognition as well as the ability for the assistant to be able to understand vernacular language. He further goes on to explain how if the virtual assistant would not be able to do this then it could lead to the users with cognitive disabilities to stop using it due to a feeling that it was pointing out and discriminating against them because of their disabilities.

E. Comparison of the Average user to those with Special Needs

The difference between what an Average user versus those with either an Intellectual Disability or who are cognitively impaired can differ greatly based on the previously discussed articles. The main commonality between the two groups

would, of course, be the fact that they use a Virtual Assistant with the hopes to make the use of their devices easier or to help them search for information. This is where many similarities end though with the second group having a more personal interactive approach to the virtual assistants. As mentioned previously something that was specifically mentioned in the article about the family given an Alexa was that they did not treat the echo as a part of the conversation or as a member of the conversation[33]. This was the exact opposite of what the participants in the study involving those who were cognitively impaired wanted when they mentioned that they wanted to be able to have a conversation and for the virtual assistant to be more interactive with them[34]. This approach to virtual assistants differs greatly to many's approaches but if we go back to Bangalore where people used their virtual assistant as a way to try to better learn English[32], would it not be extremely helpfully to have a conversational partner in their pocket with them to help in this endeavour. This contrast on how different people use virtual assistants based on their circumstances and environment helps to demonstrate and find ways to move forward with this technology that would benefit all users no matter the circumstances.

X. Security and Privacy Issues

Similar to many other components of technology, virtual assistants open the door to many potential security and privacy risks. According to the article on ACM Digital Library [39], one potential security risk relies on the popular Android built-in voice assistant module- Google Voice Search or GVS. In high-level detail, in a GVS attack, attackers take advantage of a user's phone speaker to initiate GVS commands from the background of the device. This allows attackers to prompt many different commands without the user even knowing. Commands can range from forging text and email messages, accessing a user's private information, and transmitting sensitive data and achieving remote control without any permission.

In greater detail, a GVS attack is triggered by an Android-device malware called 'VoicEmployer.' VoicEmployer can trigger GVS by playing an audio file in the background of the device where it will be undetectable by both users and the system. Due to the nature of GVS, which is a trust system built-in module, access to it from undetected audio-files could potentially lead to many possible information leaks and system prompts.

More background information on the other technologies used to perform a GVS attack reveals that the Google Services Framework and Android Intent mechanism contributes to much of the vulnerabilities encountered during an attack. The Google Services Framework comes preinstalled in almost every new Android device; this framework provides a bundle of very popular apps developed by Google including the Google Search app which includes the Google Voice Search module. The Android Intent mechanism is what allows apps to start activities or services in another app. More generally, an intent itself is a messaging object which knows of both a recipient and contains data. This intent mechanism is precisely what is used by VoicEmployer to execute malicious operations which are perceived as mere normal, in-app operations by the Android OS.

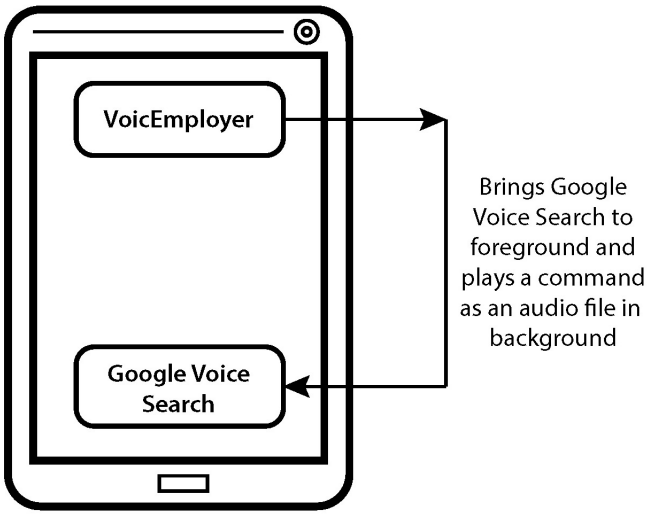

Fig. 4. Inter-Application Communication Channel of a GVS Attack - [39] VoicEmployer is an Android Intent mechanism which allows Google Voice Search to be brought into the foreground to listen to an infiltrator's commands.

A closer analysis of the Google Search app reveals the exact sequence of events that take place during the activation of the Google Voice Search module and gives a better idea of how the vulnerabilities of the Google are taken advantage of. The voice search module can run in two different modes: 'Voice Dialer' and 'Velvet' mode. Voice Dialer mode can be seen as the more watered down version of the voice search module; this mode can only accept voice dialing commands. Velvet mode is the fully functional mode which allows access to most apps and actions. Velvet mode can only be accessed when the phone is unlocked, and the screen is on, while Voice Dialer mode is activated when the phone is locked.

XI. Conclusion

Turing tests were once believed to be difficult to pass by the old technology standards. With recent advancements, in technology (such as duplex) that perfectly imitates the "umms" and "ahs" in our natural language, you may not even need the teleprinter. Not only are our machines "thinking" by Turing standards, but our machines are becoming better at imitating us. From Eliza to Google Assistant we have advanced our "thinking machines." Despite these advances, they are only intelligent for purpose-specific commands, making them still unable to reach general intelligence. Chatbots are conversational assistants inspired by the Turing test itself, and virtual assistants are newly developed purpose specific chatbots.

The framework we've developed our chatbots and virtual assistants have dramatically changed from simple input/output responses for typed answers to deep learning techniques, and natural language voiced responses. Virtual assistants are useful to assist people struggling with literacy or disabilities as well as acting as a convenience for those who would use it for general purposes.

We expect in the future voice assistants will improve upon current faults and difficulties, such as voice recognition and natural language understanding, may apply emotion theory to simulate emotion, hybrid the technology into developing technologies (such as self-driving cars) but we also expect to see some privacy/security/ethical issues surface.

REFERENCES

[1] Y. Ishida and R. Chiba, "Free will and turing test with multiple agents: An example of chatbot design," *Procedia computer science*, vol. 112, pp. 2506–2518, 2017.

[2] N. M. Radziwill and M. C. Benton, "Evaluating quality of chatbots and intelligent conversational agents," *arXiv preprint arXiv:1704.04579*, 2017.

[3] A. Deshpande, A. Shahane, D. Gadre, M. Deshpande, and P. M. Joshi, "A survey of various chatbot implementation techniques," *International Journal of Computer Engineering and Applications*, vol. 11, 2017.

[4] P. Doss, A. Pal, K. J. S. Paul, and R. SRM IST, "Unified voice assistant and iot interface," *International Journal of Engineering Science*, vol. 19061, 2018.

[5] R. Dale, "The return of the chatbots," *Natural Language Engineering*, vol. 22, no. 5, pp. 811–817, 2016.

[6] Y.-N. Chen, A. Celikyilmaz, and D. Hakkani-Tür, "Deep learning for dialogue systems," *Proceedings of ACL 2017, Tutorial Abstracts*, pp. 8–14, 2017.

[7] J. Gao, M. Galley, and L. Li, "Neural approaches to conversational ai," *arXiv preprint arXiv:1809.08267*, 2018.

[8] H. Liu, T. Lin, H. Sun, W. Lin, C.-W. Chang, T. Zhong, and A. Rudnicky, "Rubystar: A non-task-oriented mixture model dialog system," *arXiv preprint arXiv:1711.02781*, 2017.

[9] F. Strub, H. De Vries, J. Mary, B. Piot, A. Courville, and O. Pietquin, "End-to-end optimization of goal-driven and visually grounded dialogue systems," *arXiv preprint arXiv:1703.05423*, 2017.

[10] G. Lugano, "Virtual assistants and self-driving cars," in *2017 15th International Conference on ITS Telecommunications (ITST)*. IEEE, 2017, pp. 1–5.

[11] X. Rong, A. Fourney, R. N. Brewer, M. R. Morris, and P. N. Bennett, "Managing uncertainty in time expressions for virtual assistants," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2017, pp. 568–579.

[12] R. Sarikaya, "The technology behind personal digital assistants: An overview of the system architecture and key components," *IEEE Signal Processing Magazine*, vol. 34, no. 1, pp. 67–81, 2017.

[13] S. Quarteroni, "Natural language processing for industry," *Informatik-Spektrum*, vol. 41, no. 2, pp. 105–112, 2018.

[14] B. Liu and I. Lane, "An end-to-end trainable neural network model with belief tracking for task-oriented dialog," *arXiv preprint arXiv:1708.05956*, 2017.

[15] I. V. Serban, R. Lowe, P. Henderson, L. Charlin, and J. Pineau, "A survey of available corpora for building data-driven dialogue systems," *CoRR*, vol. abs/1512.05742, 2015. [Online]. Available: http://arxiv.org/abs/1512.05742

[16] A. Sriram, H. Jun, Y. Gaur, and S. Satheesh, "Robust speech recognition using generative adversarial networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5639–5643.

[17] R. W. White, "Skill discovery in virtual assistants," *Communications of the ACM*, vol. 61, no. 11, pp. 106–113, 2018.

[18] B. Liu, G. Tur, D. Hakkani-Tur, P. Shah, and L. Heck, "Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems," *arXiv preprint arXiv:1804.06512*, 2018.

[19] L. E. Asri, H. Schulz, S. Sharma, J. Zumer, J. Harris, E. Fine, R. Mehrotra, and K. Suleman, "Frames: A corpus for adding memory to goal-oriented dialogue systems," *arXiv preprint arXiv:1704.00057*, 2017.

[20] M. Eric and C. D. Manning, "Key-value retrieval networks for task-oriented dialogue," *arXiv preprint arXiv:1705.05414*, 2017.

[21] J. Li, W. Monroe, T. Shi, S. Jean, A. Ritter, and D. Jurafsky, "Adversarial learning for neural dialogue generation," *arXiv preprint arXiv:1701.06547*, 2017.

[22] H. Schulz, J. Zumer, L. E. Asri, and S. Sharma, "A frame tracking model for memory-enhanced dialogue systems," *arXiv preprint arXiv:1706.01690*, 2017.

[23] A. Bordes, Y.-L. Boureau, and J. Weston, "Learning end-to-end goal-oriented dialog," *arXiv preprint arXiv:1605.07683*, 2016.

[24] X. Li, Y.-N. Chen, L. Li, J. Gao, and A. Celikyilmaz, "End-to-end task-completion neural dialogue systems," *arXiv preprint arXiv:1703.01008*, 2017.

[25] T.-H. Wen, D. Vandyke, N. Mrksic, M. Gasic, L. M. Rojas-Barahona, P.-H. Su, S. Ultes, and S. Young, "A network-based end-to-end trainable task-oriented dialogue system," *arXiv preprint arXiv:1604.04562*, 2016.

[26] V. kepuska and G. Bohouta, "Next-generation of virtual personal assistants (microsoft cortana, apple siri, amazon alexa and google home)," in *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 2018, pp. 99–103.

[27] H. De Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. Courville, "Guesswhat?! visual object discovery through multi-modal dialogue," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5503–5512.

[28] V. Chattaraman, W.-S. Kwon, J. E. Gilbert, and K. Ross, "Should ai-based, conversational digital assistants employ social- or task-oriented interaction style? a task-competency and reciprocity perspective for older adults," *Computers in Human Behavior*, vol. 90, pp. 315 – 330, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0747563218304230

[29] Z. Wang, N. Cheng, Y. Fan, J. Liu, and C. Zhu, "Construction of virtual assistant based on basic emotions theory," vol. 3784 LNCS, Beijing, China, 2005, pp. 574 – 581. [Online]. Available: http://dx.doi.org/10.1007/11573548_74

[30] G. Campagna, R. Ramesh, S. Xu, M. Fischer, and M. S. Lam, "Almond: The architecture of an open, crowdsourced, privacy-preserving, programmable virtual assistant," Perth, WA, Australia, 2017, pp. 341 – 350, high-level domain;Internet of thing (IOT);Multiphttps://www.overleaf.com/project/5ca57171e166ef203a218746le devices;Natural language interfaces;Natural languages;Privacy preserving;Public knowledge;Virtual assistants;. [Online]. Available: http://dx.doi.org/10.1145/3038912.3052562

[31] "What is interoperability?" Apr 2019. [Online]. Available: https://www.himss.org/library/interoperability-standards/what-is-interoperability

[32] A. Bhalla, "An exploratory study understanding the appropriated use of voice-based search and assistants," in *Proceedings of the 9th Indian Conference on Human Computer Interaction*, ser. IndiaHCI'18. New York, NY, USA: ACM, 2018, pp. 90–94. [Online]. Available: http://doi.acm.org.lib-proxy.fullerton.edu/10.1145/3297121.3297136

[33] M. Porcheron, J. E. Fischer, S. Reeves, and S. Sharples, "Voice interfaces in everyday life," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ser. CHI '18. New York, NY, USA: ACM, 2018, pp. 640:1–640:12. [Online]. Available: http://doi.acm.org.lib-proxy.fullerton.edu/10.1145/3173574.3174214

[34] M. Baldauf, R. Bösch, C. Frei, F. Hautle, and M. Jenny, "Exploring requirements and opportunities of conversational user interfaces for the cognitively impaired," in *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*, ser. MobileHCI '18. New York, NY, USA: ACM, 2018, pp. 119–126. [Online]. Available: http://doi.acm.org.lib-proxy.fullerton.edu/10.1145/3236112.3236128

[35] J. Sanders and A. Martin-Hammond, "Exploring autonomy in the design of an intelligent health assistant for older adults," in *Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion*, ser. IUI '19. New York, NY, USA: ACM, 2019, pp. 95–96. [Online]. Available: http://doi.acm.org.lib-proxy.fullerton.edu/10.1145/3308557.3308713

[36] "Definition of intellectual disability." [Online]. Available: http://aaidd.org/intellectual-disability/definition

[37] S. S. Balasuriya, L. Sitbon, A. A. Bayor, M. Hoogstrate, and M. Brereton, "Use of voice activated interfaces by people with intellectual disability," in *Proceedings of the 30th Australian Conference on Computer-Human Interaction*, ser. OzCHI '18. New York, NY, USA: ACM, 2018, pp. 102–112. [Online]. Available: http://doi.acm.org.lib-proxy.fullerton.edu/10.1145/3292147.3292161

[38] C. for Disease Control, Prevention *et al.*, "Cognitive impairment: A call for action, now," *CDC, Atlanta, GA*, 2011. [Online]. Available: https://www.cdc.gov/aging/pdf/cognitive_impairment/cogImp_ca_final.pdf

[39] W. Diao, X. Liu, Z. Zhou, and K. Zhang, "Your voice assistant is mine: How to abuse speakers to steal information and control your phone," in *Proceedings of the 4th ACM Workshop on Security and Privacy in Smartphones & Mobile Devices*, ser. SPSM '14. New York, NY, USA: ACM, 2014, pp. 63–74. [Online]. Available: http://doi.acm.org.lib-proxy.fullerton.edu/10.1145/2666620.2666623