1 Acknowledgement

No volume of words is enough to express my gratitude towards my guide, Prof. A .M. Bhadgale, Assistant Professor in Computer Engineering Department, who has been very concerned and have aided for all the help essential for the preparation of this work. He has helped me to explore this vast topic in an organized manner and provided me with all the ideas on how to work towards a research-oriented venture.

I am also thankful to Prof. M. V. Marathe, Head of Department, Computer Engineering, for the motivation and inspiration that triggered me for the seminar work.

Akash Rasal T150074255

(T.E. Computer Engineering)

2 Abstract

One of the goals of Artificial intelligence (AI) is the realization of natural dialogue between humans and machines. in recent years, the dialogue systems, also known as interactive conversational systems are the fastest growing area in AI. Many companies have used the dialogue systems technology to establish various kinds of Virtual Personal Assistants(VPAs) based on their applications and areas, such as Microsofts Cortana, Apples Siri, Amazon Alexa, Google Assistant, and Facebooks M. However, in this proposal, we have used the multi-modal dialogue systems which process two or more combined user input modes, such as speech, image, video, touch, manual gestures, gaze, and head and body movement in order to design the Next- Generation of VPAs model. The new model of VPAs will be used to increase the interaction between humans and the machines by using different technologies, such as gesture recognition, image/video recognition, speech recognition, the vast dialogue and conversational knowledge base, and the general knowledge base. Moreover, the new VPAs system can be used in other different areas of applications, including education assistance, medical assistance, robotics and vehicles, disabilities systems, home automation, and security access control.

Contents

3 Holobot

Holobot is a combination of hologram and virtual personal assistant devices. Holograms on integration with virtual personal assistant will make an exceptionally well duo. It will combine a complex physics and a complex computer related technology to form a new device. Holobots will be enabled with cameras, projectors, kinnect, etc. Gesture recognition, emotion recognition, holographic videochats will become possible with holobots.

3.1 Gesture recognition

Gesture recognition is a type of perceptual computing user interface that allows computers to capture and interpret human gestures as commands. It's gemeral defination can be stated as computer's ability to understand gestures and execute comands based on those gestures. How is a gesture defined? To understand how gesture recognition works, we should understand how the word gesture is defined. In its general sense, the word gesture can be refered as any non-verbal communication that is intended to communicate a specific message. In the world of gesture recognition, a gesture is defined as any physical movement, large or small, that can be interpreted by a motion sensor. It may include anything from the pointing of a finger to a roundhouse kick or a nod of the head to a pinch or wave of the hand. Gestures can be broad or small and contained. Even voice or verbal commands may be considered as gestures in some cases. Kinect looks at a range of human characteristics to provide the best command recognition based on natural human inputs. It provides skeletal and facial tracking along with gesture recognition, voice recognition and in some cases the depth and color of the background scene. Kinect reconstructs all of this data into printable three-dimensional (3D) models. The latest Kinect developments include an adaptive user interface that can detect a users height.

Figure 1: Holobot demo by Microsoft

3.2 Emotion Recognition

Emotion recognition is used in softwares that allows a program to read the emotions on a human face using advanced image processing. Companies are experimenting with combining sophisticated algorithms with image processing techniques that have emerged in the past ten years to understand more about what an image or a video of a person's face tells us about how he/she is feeling. Advances in technology have enabled emotion recognition. Software has become very capable. Besides its ability to track basic facial expressions for emotion such as sadness, happiness, anger, surprise, etc., these softwares can also capture "micro-expressions" or subtle body language cues that may betray an individuals feelings without his/her knowledge. In many senses, emotion recognition goes along with other kinds of facial profiling such as biometric image recognition. And both of these types of technology can be used for the same types of security purposes. Law enforcers can use emotion recognition software to try to get more information about someone during interrogation. Emotion recognition continues to evolve in much the same way as other new technologies like natural language processing are advancing, and these advances are largely made possible by the combination of ever more powerful processors, the scientific development of sophisticated algorithms, and other related innovations.

Human voice changes slightly according to their mood. These slight changes in voice and facial recognition can be used to correctly detect mood of the person. This can be then used to comfort the user accordingly and his mood may be changed. Virtual asistants enabled with gesture and emotion recognition will take VPA's a step further.

3.3 Holographic video chats

Microsofts holoportation demo shows that its possible to have holographic video chats- if youve couple of Hololenses and a couple of rooms surrounded by 3D cameras anyway. Its not a perfect re-creation: you can tell that youre looking at a hologram. But its not hard to imagine this being more intimate than a Skype chat. The setup isnt simple. A HoloLens is required to see the holograms in realtime, and a room surrounded by 3D cameras is necessary to create them. Hololens is required to see other people that are holographically projected. You cant make eye contact with people who are wearing the augmented reality device. To see amother person's face they need to take their hololens off which will make them unable to see the other user.

In one of the demonstrations by Microsoft, Izadi Interacts with his daughter Lilly, who is not wearing a HoloLens. His daughter cannot see him, and almost walks straight through him because of this. Theres a design problem here that headsets cant readily solve. But its a remarkable example of what sort of tech well have access to in the future, and its particularly cool to see the recorded conversation played back, and even shrunk.

The recent hologram call was limited to a 3D image on a monitor. But its an exciting development that suggests holographic calling isnt too far away though, well at least have to wait until 5G networks arrive in full. Most estimates suggest 2020 as the year for an initial rollout of 5G network technology, but there has to be an agreed standard before that happens. Itll be a while before we see holographic calls arriving, but work has already started on making them a reality. KT is working on the holographic call as a set of 5G-based immersive media which also includes 360-degree Live VR.

VPA's which have holograms included in their hardware will enhance our communication experience. Holograms in near future will become an integral part of our life. Also, visualisation of many complex designs can be seen unsing holograms embeded in the virtual personal assistants.

4 VPAs for commercial offices

Limiting virtual personal assistants to homes will be restricting its abilities. VPAs can be used in office to read out emails, setup meeting, and see attachments. Commercial VPAs will be enabled with cameras, projectors, etc. Cameras can be used for gesture recognition while projectors can be used to view attachments of the emails. Also, text summarisation can be done on the emails to hear exact summary instead of hearing the whole email.

4.1 Text summarisation

What Is Automatic Text Summarization? Summarizer is a micro service which uses the Classifier4J framework and its summarization module to scan through large documents and returns the sentences most likely useful for generating a summary. Automatic summarization works by first calculating the word frequencies for the entire text document. Then, the most common words are stored and sorted. Each sentence is then scored based on how many high frequency words it contains, with higher frequency words being worth more. Finally, top X sentences are taken, and sorted based on their position in the original text. The automatic text summarization algorithm is able to function in a variety of situations that other implementations might struggle with, such as documents containing foreign languages or unique word associations that arent found in Standard English language corpuses.

Why You Need Text Summarization in VPA's? Business leaders, analysts, paralegals, and academic researchers need to comb through huge numbers of documents every day to keep ahead, and a large portion of their time is spent just figuring out what document is relevant and what isnt. By extracting important sentences and creating comprehensive summaries, it is possible to quickly assess whether or not a document is worth reading. Automatic text summarization is also useful for students, authors and poets. Being able to automatically generate an absjustract based for your research paper or chapter in a book in a clear and concise way that is faithful to the original source material!

Abstraction-based summarization

The abstraction technique entails paraphrasing and shortening parts of the source document. When abstraction algorithm is applied for text summarization in deep learning problems, it can overcome the grammar inconsistencies of the extractive method. The abstractive text summarization algorithms create new sentences that relay the most useful information from the original text just like humans do. However, the text summarization algorithms required to do abstraction are more difficult to develop; thats why another ssumarisation technique, extraction is still popular. Here is an example: Abstractive summary: Joseph and Mary came to Jerusalem where Jesus was born. How does a text summarization algorithm work? Usually, text summarization in NLP is treated as a supervised machine learning problem (where future outcomes are predicted based on provided data). Typically, here is how using the extraction-based approach to summarize texts can work: 1. Introduce a method to extract the merited keyphrases from the source document. For example, you can use part-of-speech tagging, words sequences, or other linguistic patterns to identify the keyphrases. 2. Gather text documents with positively-labeled keyphrases. The keyphrases should be compatible to the stipulated extraction technique. To increase accuracy, you can also create negatively-labeled keyphrases. 3. Train a binary machine learning classifier to make the text summarization. Some of the features you can use include: Length of the keyphrase Frequency of the keyphrase The most recurring word in the keyphrase Number of characters in the keyphrase 4. Finally, in the test phrase, create all the keyphrase words and sentences and carry out classification for them.

5 Smart displays

What is a smart display?

Smart displays add a touchscreen to the mix, which lets you watch videos or look at pictures. The smart displays we've tested can also walk you through a recipe step by step, show you detailed weather forecasts or give info about restaurants if you're searching for something to eat.

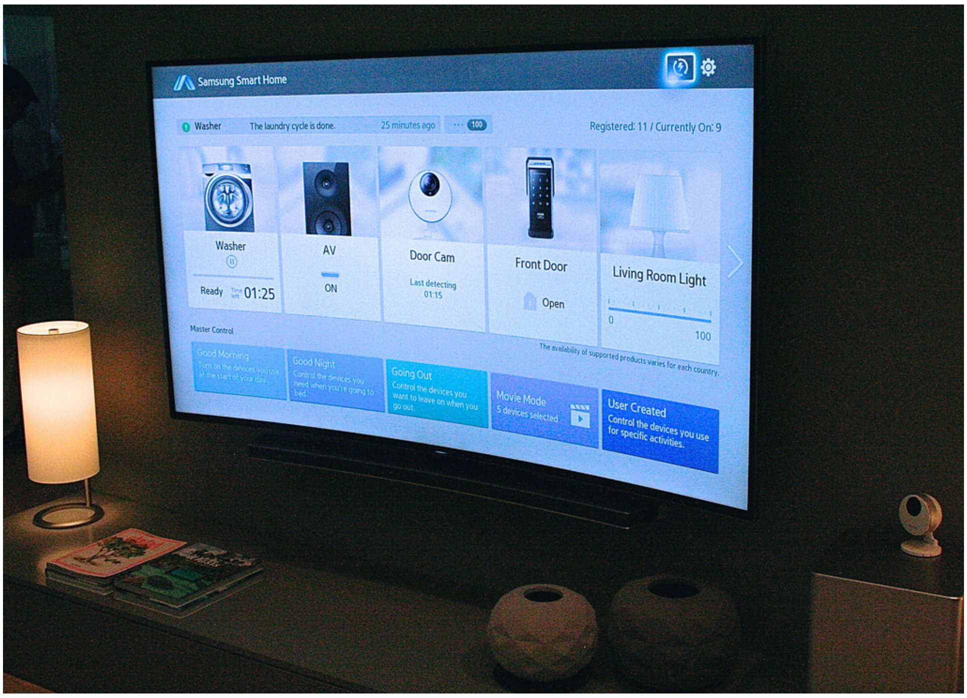


Figure 2: Concept Smart Screen

The second-gen Amazon Echo Show ushered in an updated wave of smart displays. Tyler Lizenby/CNET. Smart displays either use Amazon's digital assistant Alexa or Google's competitive Google Assistant. Options with Alexa are the newly released second-generation Amazon Echo Show and the upcoming Facebook Portal. Google Assistant is built for the Lenovo Smart Display, the JBL Link View and the upcoming smart displays from LG and Sony. The upcoming Google Home Hub also features Google Assistant.

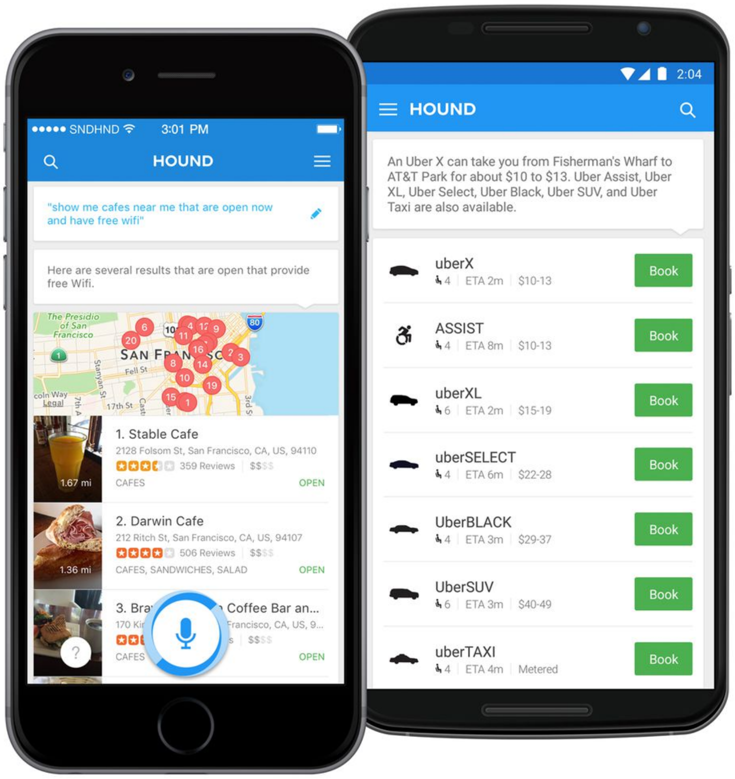
Screen sizes ranges from 7 to 10 inches, so smart displays won't replace your television. Without traditional apps, they are also not as functional as tablets, though you will find the same access to Google's Actions and Alexa's Skills, which are voice-centric apps. Smart displays aren't really designed for surfing the web. They offer a simple interface you can see from across the room and are for simple interactions designed to enhance your initial voice query.

All smart displays also allow you to make video calls, but the upcoming Google Home Hub is an outlier in that it doesn't have a camera. You can still make video calls with it, but the person you're calling won't be able to see you (and yes, some folks might prefer that approach). The rest smart displays all have cameras for making two-way video calls.

Upcoming smart displays may have algorithms to identify individuals in the house. It can specifically show related contents to the individual instead of broadcasting it. They may review our clothes, suggest fashion and even visualise ourselves in different get-ups. They can display current temperature and weather predictions.

6 Hound

HOUND Voice Search and Mobile Assistant content rating is everyone. This app is listed in Productivity category of app store. You may visit SoundHound Inc.'s website to know more about the company/developer who developed this. HOUND Voice Search and Mobile Assistant can be downloaded on android devices supporting 16 api and above. Download the app using your app store and click on install to install the app. You could also download apk of HOUND Voice Search and Mobile Assistant and run it using popular android emulators.



"show me cafes near me that are open now and have free wifi"

Here are several results that are open that provide free Wifi.

1. Stable Cafe
2128 Folsom St, San Francisco, CA, US, 94110
359 Reviews | $$$$
CAFES OPEN
1.67 mi

2. Darwin Cafe
212 Ritch St, San Francisco, CA, US, 94107
506 Reviews | $$$$
CAFES, SANDWICHES, SALAD OPEN
1.36 mi

3. Bra… Coffee Bar an…
170 Ki… Francisco, CA, US, 9…
$$$$
CAFES OPEN

An Uber X can take you from Fisherman's Wharf to AT&T Park for about $10 to $13. Uber Assist, Uber XL, Uber Select, Uber Black, Uber SUV, and Uber Taxi are also available.

uberX	4 | ETA 2m	$10-13	Book
ASSIST	4 | ETA 8m	$10-13	Book
uberXL	6 | ETA 2m	$15-19	Book
uberSELECT	4 | ETA 6m	$22-28	Book
UberBLACK	4 | ETA 3m	$29-37	Book
UberSUV	6 | ETA 3m	$40-49	Book
uberTAXI	4 | ETA 4m	Metered	Book

Figure 3: Hound app display

Nearly every large technology companies is trying to figure out how to let computers understand human speech, but a Santa Clara-based startup have just cut its way to the top of the field. Hound, as its app is called, is a voice-powered digital assistant. You can talk to it, ask questions, and have it perform tasks for you. What sets Hound apart is its faster and more capable than anything you've ever tried before. SoundHound, the company behind Hound, launched the app in beta for Android users in 2016 Summer, and in eight months improving the service with the help of about 150,000 testers. The company has also launched partnerships with Yelp and Uber today to let Hound users get restaurant information and hail a ride from within the app. Those integrations are nice, but Hound has a tall order: it's trying to usurp the likes of Apple and Google as the go-to voice interface for smartphones. Talking to our technology is considered one of the next big leaps in computing. When software has the ability to understand what we're saying and how we're saying it, it'll be able to parse questions and supply answers, perform tasks on our behalf, and transform how we interact with devices. So far that vision hasn't quite arrived. Apple's Siri often stumbles on simple requests, while Google Now is a personality-devoid arm of the company's search engine. Microsoft's Cortana is trying to be both clever and useful, but it's virtually nonexistent on mobile phones where we need it most. Amazon's Alexa is gaining steam in the smart home, but you can't ask it anything complex.

Hound is the first digital assistant that feels like a real step toward the future, albeit a handicapped one at the moment. It's not that Hound feels more like you're talking to a human it's quite robotic in fact but it is without a doubt the smartest and fastest voice-based assistant I've ever seen. The app is extremely fast that it can produce near real-time translations of whole sentences in other languages, and it can spit back mounds of requested data faster than you could ever possibly glean it from Google with a keyboard. You have to open the app to ask it questions, which is a drag. The company has added 3D Touch support to its iOS version so you can jump right into a query and a ”Ok Hound” voice command to make hands-free requests. The software's true appeal is understanding questions within questions and sussing out human context. You can give it sprawling, absurd requests nested inside other requests like, ”What is the population and capitals of Japan and China, their area in square miles, and the population of India, and the area codes of France, Germany, and Spain?” and Hound will give you the information just seconds later. It remembers too, allowing you to try follow-up questions. Ask Hound to find you a coffee shop within walking distance that has free Wi-Fi, and you can then tell it to exclude Starbucks to narrow the search. You can ask it for hotels costing between $200and400$ a night in Seattle near the Space Needle and when the sun will rise two days before Christmas four years from now. All of it feels hyper-specific, and Hound's default screen is a help section guiding users on what the service is even capable of. But its underlying power is undeniably impressive.

SoundHound CEO Keyvan Mohajer won't disclose how the company's software is able to do this when Apple and Google cannot. He says the underlying technology behind Hound is built around a unique approach to natural language processing. When combined with advances in machine learning and other artificial technology techniques, Hound is able to do what Mohajer calls ”speech-to-meaning.” While other digital assistant software translates what you speak into text and tries to figure out what you said, Hound supposedly skips that step and deciphers your speech as it hears it. ”Weve been working on it for nine years. Its not a new direction,” Mohajer tells me. ”It was an original ambition of the company and we knew it was going to take that long.” SoundHound has been spent the better part of the last decade as a lesser-known Shazam. The company's main mobile app identifies music and differentiated itself early on by letting you hum a tune into your phone to hear the song and artist name. SoundHound's proficiency at audio recognition has helped it license out its technology to businesses. Yet all the while Mohajer says the startup has been preparing for the moment when a digital assistant could make use of modern software advances and surpass well-known competitors.

It's not perfect. When Hound does stumble, it does so in weird, inconsistent ways. SoundHound built its own so-called knowledge graph it doesn't use Google from which the app pulls information. Sometimes that information simply isn't there. It can tell us when President Obama's grandmother was born, but can't tell me who won the Oscar for best supporting actor in 2003. It also defaults to Microsoft's Bing for anything it doesn't understand. More often than not if it's kicking you to search it's because it either misunderstood what you were asking or it just didn't know how to answer. The goal of Hound is to avoid doing that so often that you ditch using the app altogether. Some of those inconsistencies will undoubtedly be smoothed over with enough time. But there will always be things Hound can't do because it doesn't have direct access to your phone's software or intimate knowledge of your search history and email, like Google. Right now, Hound is limited to what it's best at. For the most part, that includes finding local businesses, asking oddball queries on the fly, and doing language translation. It can also tell you how much an Uber ride will cost without having to input your pickup location and destination in the Uber app, which is a neat benefit I can see myself using all the time.

Its CEO says the company has created an easy way to develop new domains, which is industry jargon for a specific skill set aided by a third-party API. For instance, there's a weather domain, a hotels domain provided by Expedia, and a phone domain that each let Hound take your question or request and execute it. SoundHound says it started out in beta with around 50 domains and has grown to more than 100. Mohajer says most of the company's competitors don't offer much more than two dozen simple domains.

Still, it's hard not to think SoundHound's technology could be so much more useful and widespread if it were adopted by giants like Apple or Google, which would require an acquisition that may never happen. For now, SoundHound is intent on making its product the best alternative out there, and it hopes other companies will rally behind the product if it remains one step ahead of Siri and others. "I use Google Maps on iOS instead of Apple Maps, even though Apple Maps is more integrated," Mohajer told The Verge when Hound first launched in beta. "I think if you deliver something that is substantially better, people will use it." Hound isn't quite there yet, but it is most certainly on its way.

7 SMART-ER homes

SMART home technology use devices connected to the Internet of things (IoT) to automate and monitor in-home systems. SMART stands for Self-Monitoring Analysis and Reporting Technology. The technology was first developed by IBM and was referred to as Predictive failure analysis. The first contemporary SMART home technology products became available to consumers between 1998 and the early 2000s. SMART home allows users to control and monitor their connected home devices from SMART home apps, smartphones, or other networked devices. Users can remotely control connected home systems whether they are at home or away. This allows more efficient energy and electric use as well as ensuring your home is secure. SMART home technology contributes to health and well-being enhancement by accommodating people with special needs, especially older people. SMART home technology is now being used to create SMART cities. A Smart city functions similar to a SMART home, where systems are monitored to more efficiently run the cities and save money.

As of 2015, the most common piece of SMART home technology in the United States were wireless speaker systems with 17 percent of people having one or more. SMART thermostats were the second most prevalent piece of SMART home technology with 11 percent of people using the device. A 2012 consumer report that pulled data from the National Association of Home Builders looked for what SMART home devices homeowners wanted most and found that top five were wireless security systems (50 percent), programmable thermostats (47 percent), security cameras (40 percent), lighting control systems and wireless home audio systems (39 percent), and home theater and multi-zone HVAC systems (37 percent).

The small actions like pointing towards a television remote can detected by the smarter home using gesture recognition. It may bring us the remote by using some new method. Also, machine learning can be used to learn about the time required to get ready for breakfast from the time of waking up. The home can on its own make breakfast matching the predicted time of the users breakfast. Many devices need to be incorporated to perform all these futuristic actions like cameras, microphones, etc.

8 Conclusion

This proposal introduces the structure of Next-Generation of Virtual Personal Assistants that is a new VPAs system designed to converse with a human, with a coherent structure. This VPAs system has used speech, graphics, video, gestures and other modes for communication in both the input and output channel. Also, the VPAs system will be used to increase the interaction between users and the computers by using some technologies such as gesture recognition, image/video recognition, speech recognition, and the Knowledge Base. Moreover, this system can enable a lengthy conversation with users by using the vast dialogue knowledge base. Moreover, this system can be used in different tasks such as education assistance, medical assistance, robotics and vehicles, disabilities systems, home automation, and security access

control. Also, it can be a satisfactory solution that can be used by applications, such as responding to customers, customer service agent, training or education, facilitating transactions, online shopping, travelling information, counseling, tutoring system, ticket booking, remote banking, travel reservation, Information enquiry, stock transactions, taxi bookings, and route planning etc. In the end, to achieve the final stage and all these improvements to the new system with high accuracy, we need funding from an organization that will work with us to improve the system by funding the new hardware devises that have high accuracy, as well as the tools and cloud severs that we will need for testing the new system.

9 References

1. https://towardsdatascience.com/a-quick-introduction-to-text-summarization-in-machine-learning-3d27ccf18a9f

2. https://devhub.io/repos/ActiveNick-HoloBot

3. https://www.theverge.com/2016/3/1/11136298/hound-app-ios-android-siri-google-now-cortana