Next-Generation of Virtual Personal Assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home)

Veton Këpuska
Electrical & Computer Engineering Department
Florida Institute of Technology
Melbourne, FL, USA
vkepuska@fit.edu

Gamal Bohouta
Electrical & Computer Engineering Department
Florida Institute of Technology
Melbourne, FL, USA
gidris2015@my.fit.ed

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 Microsoft's Cortana, Apple's Siri, Amazon Alexa, Google Assistant, and Facebook's 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.**

Keywords— Virtual Personal Assistants; Multi-modal Dialogue Systems; Gesture Recognition; Image Recognition; Image Recognition.

I. INTRODUCTION

Spoken dialogue systems are intelligent agents that are able to help users finish tasks more efficiently via spoken interactions. Also, spoken dialogue systems are being incorporated into various devices such as smart-phones, smart TVs, in car navigating system [9]. Also, Dialogue systems or conversational systems can support a wide range of applications in business enterprises, education, government, healthcare, and entertainment. Personal assistants, known by various names such as virtual personal assistants, intelligent personal assistants, digital personal assistants, mobile assistants, or voice assistants [7].

Many companies have used the spoken dialogue systems to design their dialogue system device, such as Microsoft's Cortana, Apple's Siri, Amazon Alexa, Google Assistant, Samsung S Voice, Nuance Dragon, and Facebook's M. These companies used different approaches to design and improve their dialogue systems. There are many techniques used to design the VPAs, based on the application and its complexity. For example, Google has improved the Google Assistant by using the Deep Neural Networks (DNN) method which highlights the main components of dialogue systems and new deep learning architectures used for these components [16]. Also, Microsoft used the Microsoft Azure Machine Learning Studio with other Azure components to improve the Cortana dialogue system [11].

Moreover. The Amazon provides the advanced deep learning functionalities of automatic speech recognition (ASR) for converting speech to text, and natural language understanding (NLU) to recognize the intent of the text, to enable developer to build applications with highly engaging user experiences and lifelike conversational interactions [10]. Also, Facebook has launched its own personal assistant, Messenger M, which is working to combine machine-learning algorithms with contextual memory. Facebook is training Facebook's new virtual assistant for Messenger with supervised learning, a process where the computer learns by example from what human trainers teach it [13]. All these companies are trying to develop the competences in several of the core technologies for their dialogue systems, such as automatic speech recognition, text-to-speech, synthetic talking face and dialog management.

Fig. 1. (Apple's HomePod (a), Amazon's Echo (b), Microsoft/Harman Kardon's Cortana (c), and the Google Home (d))

Moreover, there are some companies and researchers that have attempted to improve their applications by using the Multi-modal dialogue technique to design the Next-Generation of dialogue systems. The Multi-modal dialogue process two or more combined user input modes, such as speech, pen, touch, manual gestures, gaze, and head and body movement. For example, In the Ford Model U Concept Vehicle, this system, including a touch screen and a speech recognizer, is used for controlling several non-critical automobile operations, such as climate, entertainment, navigation, and telephone. The prototype implements a natural language spoken dialog interface integrated with an intuitive graphical user interface, as opposed to the traditional, speech only, command-and-control interfaces deployed in some of the vehicles currently on the market [4].

Also, Semio is a part of research at the University of Southern California: "Semio is developing a cloud-based platform to allow humans to use robots through natural communication-speech and body language. The platform allows developers to create and deploy speech/gesture-based applications to be executed by robots and allows non-expert users to access and use those robot applications through natural communication" [2].

Also, Kuniaki Noda, Hiroaki Arie, Yuki Suga, and Tetsuya Ogata from Waseda University, proposed a deep neural network framework that enables multimodal integration learning of temporal sequences, including visual, auditory, and motion. The performance of their proposed framework was evaluated by two tasks utilizing a humanoid robot in a real-world environment [3]. Moreover, Zhou Yu from Carnegie Mellon University, designed and implemented a multimodal dialog system to coordinate with users' engagement and attention on the fly via techniques such as adaptive conversational strategies and incremental speech production. This work uses automatic extracted features in three modalities: text, audio and video [4].

In this proposal, we propose an approach that will be used to design the Next-Generation of Virtual Personal Assistants, increasing the interaction between users and the computers by using the Multi-modal dialogue system with techniques including the gesture recognition, image/video recognition, speech recognition, the vast dialogue and conversational knowledge base, and the general knowledge base. Moreover, our approach will be used in different tasks including education assistance, medical assistance, robotics and vehicles, disabilities systems, home automation, and security access control.

Also, the approach has some new methods that make this device unique, such as using as TV by using the data show or connecting the device with screen, watching TV and movies with translation language, chatting with anyone in any language, understanding body language and movements, and playing games with speech and gesture recognition; it also can be used to read facial and speech expressions. To design the Next-Generation of Virtual Personal Assistants with high accuracy, we added some components to the original structure of general dialogue systems to change the general model to Multi-modal dialogue systems, such as ASR Model, Gesture Model , Graph Model, Interaction Model, User Model, Input Model, Output Model, Inference Engine, Cloud Servers and Knowledge Base.

II. THE STRUCTURE OF GENERAL DIALOGUE SYSTEM

The dialogue system is one of an active area that many companies use to design and improve their new systems. According to CHM Research, before 2030, millions of us will be using “voice” to interact with machine, and voice-driven services will become part and parcel of smartphones, smart glasses, home hubs, kitchen equipment, TVs, games consoles, thermostats, in-car systems and apparel [8]. There are many techniques used to design the dialogue systems, based on the application and its complexity. On the basis of method used to control dialogue, a dialogue system can be classified in three categories: Finite State (or graph) based systems, Frame based system and Agent based systems [1].

Also, there are many different architectures for dialog systems. Which sets of components are included in a dialog system, and how those components divide up responsibilities differs from system to system. A dialogue system has mainly seven components: Input Decoder, Natural Language Understanding, Dialogue Manager, Domain Specific Component, Response Generator, and Output Renderer [1]. However, there are six main components in the general dialogue systems, which includes the Speech Recognition (ASR), the Spoken Language Understanding (SLU), Dialog Manager (DM), Natural Language Generation (NLG), Text to Speech Synthesis (TTS), and the knowledge base. The following is the structure of the general dialogue system:

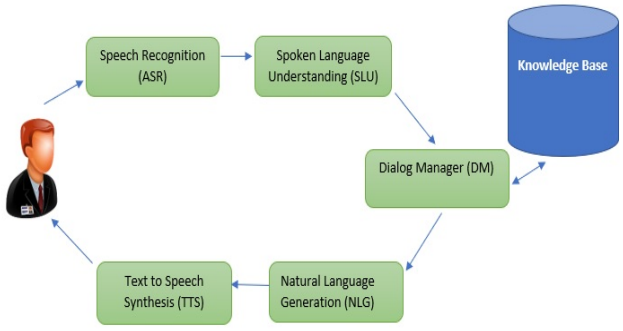

Fig. 2. The Structure of General Dialogue System

III. THE PROPOSAL VPAs SYSTEM

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. We have modified and added some components in the original structure of general dialogue systems, such as ASR Model, Gesture Model, Graph Model, Interaction Model, User Model, Input Model, Output Model, Inference Engine, Cloud Servers and Knowledge Base. The following is the structure of the Next-Generation of Virtual Personal Assistants:

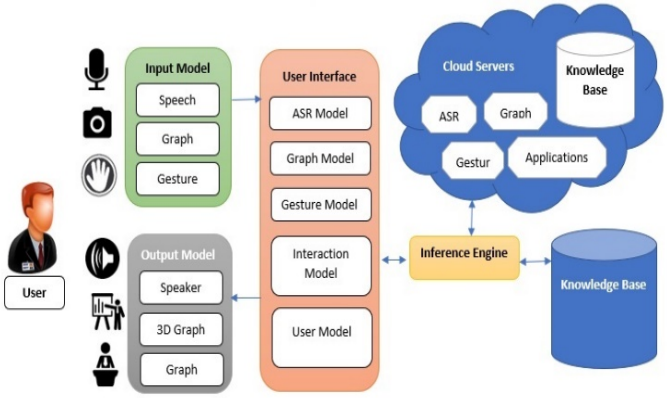

Fig. 3. The Structure of The Next-Generation of Virtual Personal Assistants

A. Knowledge Base

There are two knowledge bases. The first is the online and the second is local knowledge base which include all data and facts based on each model, such as facial and body data sets for gesture modal, speech recognition knowledge bases, dictionary and spoken dialog knowledge base for ASR modal, video and image body data sets for Graph Model, and some user's information and the setting system.

B. Graph Model

The Graph Model analyzes video and image in real-time by using the Graph Model and extracts frames of the video that collect by the camera and the input model; then it sends those frames and images to the Graph Model and applications in Cloud Servers for analyzing those frames and images and returning the result.

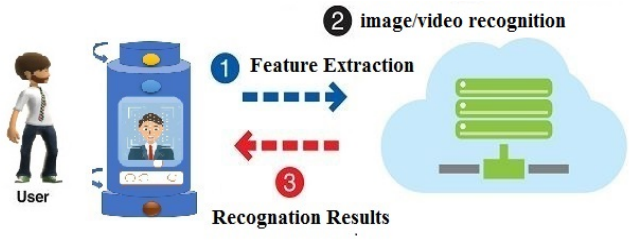

Fig. 4. The Graph Model

C. Gesture Model

The gesture model uses the camera and Kinect in the input model to read the movements of the human body and the facial expressions; then it sends all data to the gesture model and applications in Cloud Servers to analyze those frames and images and returning the result.

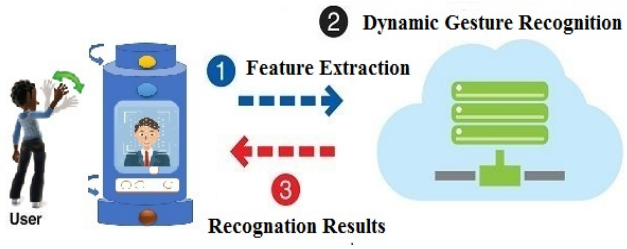

Fig. 5. The Gesture Model

D. ASR Model

The speech recognition model will work in real-time with the microphone in the input model with the ASR model in Cloud Servers to recognize the utterances that a user speaks into a microphone and then convert it to text; then it sends the text to the applications in Cloud Servers to analyze the text and returning the result.

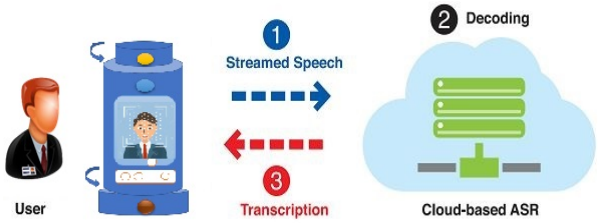

Fig. 6. The ASR Model

E. Interaction Model

This is the main model that will be used to provide interaction between users of the system and the system models by receiving the data from the input model and analyzing the data to send for each model based on its tasks, then returning result that will be used to make the final decision.

F. Inference Engine

The inference engine works together with the Interaction Model in the chain of conditions and derivations and finally deduces the outcome. they analyze all the facts and rules, then sorts them before concluding to a solution.

G. User Model

This model has all information about the users that will use the system. It can include personal information such as users' names and ages, their interests, their skills and knowledge, their goals and plans, their preferences and their dislikes or data about their behavior and their interactions with the system. All

information will be collected by asking the user some questions then storing all answers in the Knowledge Base.

H. Input Model

This model will organize the work of all input devices that the system uses to collect the different data from microphone, camera and Kinect. Also, this model includes intelligence algorithms to organize the input information before sending the data to the Interaction Model.

I. Output Model

This model will receive the final decision from the Interaction Model with an explanation, then it will choose the perfect output device to show the result such data show, speakers or screen based on the result.

IV. TESTING THE SYSTEM

The whole system will interpret the inputs from the users, it constructs queries to cloud servers and knowledge sources in order to perform tasks and retrieve information to be output by the response generation models. For testing the whole system, we separated the stages of testing the system into several stages, based on the difficulty and complexity of the new models. For example, the first stage is testing the ASR Model by using different ASR system and comparing speech recognition systems such as Microsoft API, Google API, Amazon API and CMU Sphinx, all information about this work has been published in the paper "Comparing Speech Recognition Systems (Microsoft API, Google API And CMU Sphinx)" [5]. Also, we tested the live speech translation by using the Google translation API to create the tool that will be used to watch TV with speech and text translation in real time, and to create the conversation between two users by using different languages. All information about this work has been published in the project "Live Speech Translation (Google TV)" [18]. Also, we tested the ASR Model by using the new ASR mode, Wake-Up-Word and General Speech Recognition Systems, and all information about this work have been publishing in paper "improving Wake-Up-Word and General Speech Recognition Systems" [6].

Moreover, we tested the dialogue system with different cloud severs such as Google cloud and Amazon Web Services. In the second stage, we tested the Graph Model by using some new machine learning models, such as Deep Neural Networks model and Convolutional Neural Networks. Then, we evaluated and trained the model with the testing and training data sets that we collected from different users and different environments, such as different background and locations in rooms, different user distances to the camera, and different user's characters. In the stage three, we are trying to use the Kinect sensors and camera to capture the user's facial expression and body language and tracking and analyzing non-verbal gestures. In the final stage, after testing each stage of the systems and according to the results of the models (ASR Model, Gesture Model, Graph Model, Interaction Model), we will connect all models together to test the entire system.

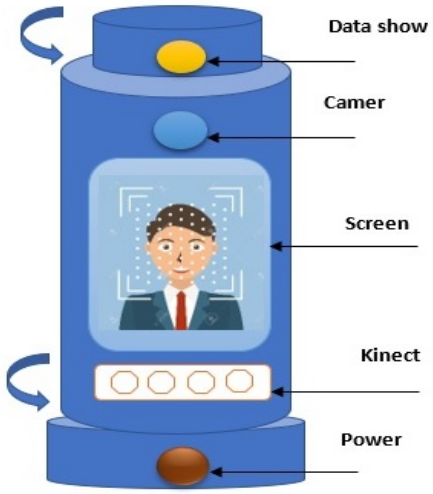

Fig. 7. The Next-Generation of Virtual Personal Assistants

V. THE EXPERIMENT RESULTS

In this proposal, we have presented the approach and tested this approach by using many tools, technologies, and speech corpus. Also, we have worked on both hardware and software sides at the same time. On the hardware side, we started our work by collecting all tools and devises that would be used in the system such as the mini data show, the mini board with CPU and LCD, the camera and Kinect. The software side, we have tested each model with different cloud severs, such as Google cloud, then we will connect all models together to test the system in the final stage. After testing the system and according to the results of the models (ASR Model, Gesture Model, Graph Model, Interaction Model), we found that the whole concept of this system is the best solution for Next-Generation of Virtual Personal Assistants by adding some improvements in the hardware and software in the final stage. 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.

VI. CONCLUSIONS

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.

REFERENCES

[1] S. Arora, K. Batra,and S. Singh. Dialogue System: A Brief Review. Punjab Technical University.

[2] R. Mead. 2017.Semio: Developing a Cloud-based Platform for Multimodal Conversational AI in Social Robotics. 2017 IEEE International Conference on Consumer Electronics (ICCE).

[3] K. Noda, H. Arie, Y. Suga,and T. Ogata. 2014.Multimodal integration learning of robot behavior using deep neural networks. Elsevier: Robotics and Autonomous Systems.

[4] R. Pieraccini, K. Dayanidhi, J. Bloom, J. Dahan, M.l Phillips. 2003. A Multimodal Conversational Interface for a Concept Vehicle. Eurospeech 2003.

[5] G. Bohouta and V. Z Këpuska. 2017. Comparing Speech Recognition Systems (Microsoft API, Google API And CMU Sphinx). Int. Journal of Engineering Research and Application 2017.

[6] G. Bohouta and V. Z Këpuska. 2017. Improving Wake-Up-Word and General Speech Recognition Systems. 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress.

[7] M. McTear .2016. The Dawn of the Conversational Interface. Springer International Publishing Switzerland 2016

[8] CM Research: High quality research requires investment 2016.

[9] Y. Nung, A. Celikyilmaz.Deep Learning for Dialogue Systems. Deep Dialogue..

[10] Amazon. Amazon Lex is a service for building conversational interfaces. https://aws.amazon.com.

[11] Microsoft. Cortana Intelligence. https://azure.microsoft.com.

[12] B. Marr. The Amazing Ways Google Uses Deep Learning AI. https://www.forbes.com..

[13] K. Wagner. Facebook's Virtual Assistant 'M' Is Super Smart. It's Also Probably a Human. https://www.recode.com.