Virtual Personal Assistant

R. Augustian Isaac
Assistant Professor, Dept. of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India.

Abishek Narayanan
B. Tech, Dept. of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India.

Abstract – Artifical Intelligence's main goal is to make Human interaction with computers and other electronic devices much easier and practical . Nowadays Personal assistants who can carry out tasks required for daily needs with just a meaningful phrase is a fast growing area . 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 , I have used a different approach to bring down the error percentage of the personal assistant. It incorporates both the visual and audio information to deduce what the person says.

Index Terms – Virtual Personal Assistant, Machine learning, Neural networks, Voice Recognition.

1. INTRODUCTION

Gone are the days when humans depended on other humans for help or services. The digitalization of the world made sure that humans no need to contact anyone else to seek help, they could depend on a far more efficient and reliable device which can take care of their everyday needs. The computers, mobiles, laptops, etc., became a part of us and our daily life, It could carry out simple calculations to complex programs to reduce monotonous work and waste of manpower.

Virtual Personal Assistant has almost become a basic necessity in all electronic devices so as to execute the required problems easily. More than just being a bot , VPA can make life easier for the user in various ways. Speech recognition is one of the relatively new integration into the VPA. But, though its moderately efficient , it is not very helpful and are not used by the user due to its high amount of error. Though the error percentage of the upcoming VPAs is around 5 percent, it still is not quite upto the mark to where it becomes a basic part of the users life. Thus the projects aim is to build a VPA with speech recognition which has a very minimal error percentage.

Voice recognition is a complex process using advanced concepts like neural networks and machine learning. The auditory input is processed and a neural network with vectors for each letter and syllable is created. This is called the data set . When a person speaks the device compares it to this vector and the different syllables are pulled out with which it has the highest correspondence. On top of this the front camera is used to capture the images of the lip movement of the user. The device is trained to recognize the words with the movements of the lips using machine learning.

2. RELATED WORK

Most of the existing projects have only used speech recognition using neural networks. Though their systems have a moderate accuracy, they are not for practical usage nor efficient to be of any real use There are a few rudimentary techniques used by them :

- Context-aware computing
- MFCC
- NLP

2.1. Context-aware computing

Context-aware computing is a class of systems that have the ability to sense their physical environment and adapt themselves to it accordingly. These can be used for recognizing words spoken by people with varying accents.It can also deduce words that may have been misspoken. The adaptation makes it very useful for real time use as most systems works only on a theoretical manner with apt conditions.

2.2. MFCC

MFCC refers to the Mel-Frequency Cepstral Coefficients. MFC (Mel-Frequency Cepstrum) is a collection of these coefficients. It amounts to the short-term power spectrum of a sound. These can be used to sense variations in sound so as to recognize the various variables required for voice recognition.

2.3. NLP

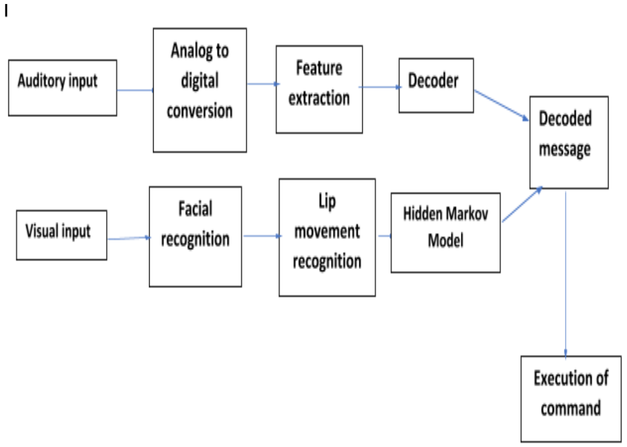
Natural Language Programming is a branch of Artificial Intelligence that deals with the interactions of computer and human languages. It mainly focuses on how to program the computers so that they can process the large lume of data on natural languages. This concept is used to familiarize the computer with the various words in a particular language and also to recognize them when spoken.

3. PROPOSED MODELLING

The major milestone that this project tries to achieve is that it tries it increase the accuracy of the speech to text software. Meaning the software will theoretically be able to convert any speech with slight modulations or different accents into text with high level of accuracy and precision needed for day to day usability of the VPA. The software essentially combines voice recognition using neural networks and lip movement detection using machine learning to increase the precision of the word spoken. For people with different accents, just voice recognition will be useless because the words they speak will be vastly different from the actual word by the computer's point of view because the vectors or the values stored for that particular word would have been gotten only based on the word being spoken in a particular accent. So here is where lip movement recognition comes into play. For most words, though in a different accent, the movement of the lips remains similar enough to deduce the word. Thus, lip movement recognition helps cutting down the various other words which would have had the same likeliness as per the voice recognition software.

4. SYSTEM ARCHITECTURE

The system architecture of this projects shows the flow of the control through the system. It also shows the hardwares and the softwares required for the execution of the program. The architecture diagram is as follows :



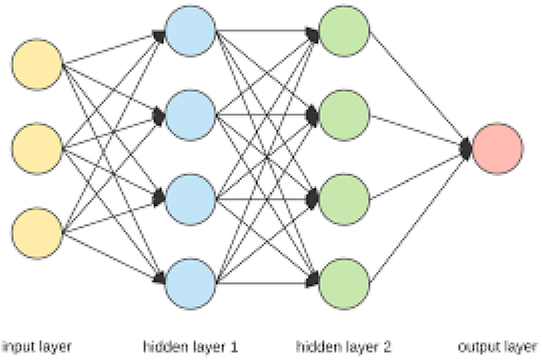
The data is collected from the camera and the auditory device then it is processed such that a meaningful word or a phrase can be deciphered. This phrase is then related to the command or response. The command is executed or the response is given as an output.

5. MODULE IDENTIFICATION

The actual working of the system is divided into four modules which is executed in order to get the end result. The modules are as follows:

- <u>VOICE INPUT</u>

The auditory sensor present in the device picks up the speech/sound emitted by the person. This sound which is stored in an analog waveform should be converted into a digital waveform. This is required for the computer to make sense of the sound and its pattern. The main concept that is being used here is neural network. Neural network is basically a part of artificial intelligence and does not have any particular tasks. When the system is provided with a large amount of data it tries to make sense or find a pattern between the data. This pattern is stored as a vector or a value with which it can later recognize the particular word or sound or phenomene. It forms various complex links over the period of training and these links can make it plausible for the system to recognize the word when spoken.



- <u>DETECTION</u>

The imagery input through the visual device gives us a chain of images of the person speaking. From these images the movement of the lips can be detected. This first requires the detection of the person's face and then hi/her lips. Then the series of detections of the lips when played gives us the movement of the lips. This movement is studied by the system and it analyzes and tries to recognize the word through a method called Hidden Markov Model. Hidden Markov Model is basically a dynamic Bayesian network, meaning the dependencies of the variables to one another gives us various outcomes depending upon the input. Hence when a phrase is said, the hidden markov model forms a acrylic graph of the various meanings that lip movement could have before it links the next word till it narrows down to the one it uniquely matches from its datasets.

- <u>DEDUCTION</u>

The system gets two similar results from the above two modules. The proportion in which they are considered depends on the way the person speaks or the reliability of the camera to capture a optimal image for the image processing. Meaning the lighting or the angle affects the proportion of the output of the module two to be considered in deciding the phrase to finalized. In the same way, the accent or the way the person pronounces the words affects the proportion of the output of module one to be considered. The algorithm used to finalize the percentage of the modules and deducing the final phrase is rather complex but is an integral part of the system that combines the voice recognition and the image processing techniques to give an efficient and accurate output.

- EXECUTION

This is the simplest part of the whole system. The output from module 3 that is the phrase spoken by the person is compared with the database to check the task that is assigned to the phrase. Meaning it is cross-referenced with the simple bot commands in the database to execute the required task given by the user.This command is then executed and the output is given on the screen.

- OUTPUT

The final output can be one of the various tasks that the personal assistant can execute. The input is basically converted into phrase which are connected to executable commands. The flow of execution takes place in the following manner: the sound is recorded and the image is taken using the camera and the microphone present in the device. This is then process using various algorithms to turn these inputs into a meaningful phrase that is present in the system's database. This phrase is then cross checked for the any commands or responses that might be connected to the phrase. Then the response is either printed or the command is executed.

6. CONCLUSION AND FUTURE WORK

This paper describes one of the most efficient ways for voice recognition. This system uses machine learning. It overcomes many of the drawbacks in the existing solutions. It is mainly built to make a much more efficient VPA o that they can be brought into much more practical day to day uses. But the system has its own limitation. Though the efficiency is high the time consumption for each task to complete maybe higher than the other VPAs and also the complexity of the algorithms and the concepts would make it very tough to tweak it if needed in the future.

ACKNOWLEDGEMENT

This project was supported by Prof. Mr.R. Augustian Isaac, hence i want to express my gratitude and thank him for giving me this opportunity to work on this project. I came across many new terms and concepts during my time in this project for which i am thankful.

REFERENCES

[1] Knote, R., Janson, A., Eigenbrod, L. and Söllner, M., 2018. The What and How of Smart Personal Assistants: Principles and Application Domains for IS Research.

[2] Feng, H., Fawaz, K. and Shin, K.G., 2017, October. Continuous authentication for voice assistants. In Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking (pp. 343-355). ACM.

[3] Canbek, N.G. and Mutlu, M.E., 2016. On the track of artificial intelligence: Learning with intelligent personal assistants. Journal of Human Sciences, 13(1), pp.592-601.

[4] Hwang, I., Jung, J., Kim, J., Shin, Y. and Seol, J.S., 2017, March. Architecture for Automatic Generation of User Interaction Guides with Intelligent Assistant. In Advanced Information Networking and Applications Workshops (WAINA), 2017 31st International Conference on (pp. 352-355). IEEE.

[5] Buck, J.W., Perugini, S. and Nguyen, T.V., 2018, January. Natural Language, Mixed-initiative Personal Assistant Agents. In Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication (p. 82). ACM.