Virtual Personal Assistant

PAHULJEET KAUR[1], P. GOWRI[2], PROF. RAJENDRA KULKARNI[3]
[1, 2, 3] *Electronics and Communication, VTU Belagavi, Guru Nanak Dev Engineering College Bidar*

Abstract- ***Artificial Intelligence (AI) technologies are one of the new technologies with new complicated features, that are emerging in a fast pace. Although these technologies seem to be extensively adopted, people do not intend to use them in some cases. Technology adoption has been studied for many years, and there are many general models in the literature describing it. However, having more customized models fore merging technologies upon their features seems necessary. In this paper, we developed a conceptual model involving a new system quality construct, i.e., interaction quality, which we believe can better describe adoption of AI-based technologies. In order to check our model, we used a voice assistant system (VAS) technology as an example of this technology, and tested a theory-based model using a data set achieved from a field survey. Our results confirm that interaction quality significantly affects individual's trust and leads to adoption of this technology.***

Indexed Terms- ***Artificial Intelligence, Voice Assistant System, Speech Recognition Model, API, Interaction Quality, Trust, Technology Adoption.***

I. INTRODUCTION

Computers can execute every command furnished by the user explicitly, even if the command assigned to the system could be in various formats. For example, commands to play music, open or print a document etc. These commands are already available in the user interface by default. The proposal for speech recognition and synthesiser has made it more straightforward rather than to use the mouse pointer to implement these commands. Speech synthesiser is the means of generating spoken language by the machine based on the written input. The ability of machines or programs to identify or acknowledge words and phrases from spoken language and convert them to a machine-readable format is known as Speech recognition. With the help of these technologies, spoken commands are performed and executed. Therefore, to make the computer recognise commands, the speech recogniser was created. As an emerging technology, most developers are not familiar with speech recognition technology; While quintessential functions of both speech recognition, and Speech synthesiser take only a few minutes to understand. Even though there are subtle and more powerful capabilities provided by computerised speech that developers shall want to understand and utilise.

II. RELATED WORK

The most famous speech recognition techniques which are existing in the real world called Cortana, Siri, and Google now. Technologies like speech recognition provide a wide variety of applications in their domain. For instance, Siri and Google now are designed particularly for the mobile phone to execute tasks like setting memo, checking messages and play a song. In contrast, Cortana is used in the computer to dictate and edit text. These commands assist the user to simulate the computer without any physical activity. It also has some commands like "Open", "Switch to "are more like natural language control, although implementation of this approach solicits the help of artificial intelligence.

Blind Source Extraction (BSE) is an approach to establish the noisy multichannel data. The preprocessing step for the speech recognition system is necessary before the BSE process. During this work, the implementation of Blind Source Extraction architecture necessitates and requires an extension of each system block within the framework for its flexibility and degree of blindness. The output of the enhancement algorithm amalgamated with the robust speech Recognition systems supported by gamma frequencies features, which are then analysed and on uncertainty, decoding to enhance the performance. Results are from different front-end, and back-end configurations manifest the benefits of these approaches.

Swamy et al [7] outlines the architectures for Automatic Speech Recognition and Voice Activity Detection in digital circuit with exceptionally enhanced accuracy, programmability, and scalability. The primary motivation for automatic speech recognition is the high requirement for memory to consume high supply. A SIMD processor with 32 parallel execution unit efficiently appraises feed-forward Deep Neural Networks for Automatic Speech Recognition. It also narrows memory consumption with a less quantised weight matrix format.

Diplophonia is a variety voice of pathological which produces the same type of two frequencies. Specialised voice analysers are used to handle up to two frequency in diplophonic voices in the earlier stages. The proposed system obtains two frequencies in diplophonic voices by using Audio Waveform Modeling by the repetitive implementation of the Viterbi algorithm. The later then executes the waveform Fourier synthesis. The variant frequencies are difficult to identify due to the fastest relevant benchmark is quite high and the average error rate in tracking 9.52%. Furthermore, illustrative results connect the speech analysis submitted.

W. Shih et al[8] represents an efficient Very Large Scale Integration Design implementation of an online repetitive processor for realtime multichannel EEG signal separation. The proposed design describes a system control unit, a single value decomposition unit, a floating matrix multiply and weight training unit. The view of the processor is varied and mixed architecture, and it differentiates the hardware parallelism as per the processing units concerning the complexity. The shared arithmetic unit and the register can significantly reduce both the complexity and power consumption in the system. The proposed solution is to use CMOS technology with 8-channel EEG processing in 128Hz rate of information, and it consumes 2.827 mW at 50 MHz clock rate. The realtime Multichannel EEG signal separation yield the highest performance as in the proposed design.

The powerful Deep Neural Networks technique is applied to incorporate the speech and produce it to waveform production artificially. The automated system speech quality is low compared to natural speech.

A Generative Adversarial Networks consist of two neural networks named discriminator and generator. A discriminator network differentiates natural and generated speeches, whereas a generator deceives the discriminator. The proposed framework includes the Generative Adversarial Networks, and the discriminator is trained with samples to discriminate the natural and the generated samples. The acoustic models; trained to reduce the sum of the traditional generation loss to the minimum and a loss for deceiving the discriminator contrasts the later. An investigation was done to find the effect of assorted Generative Adversarial Networks to the distortion and located that a Wasserstein Generative Adversarial Networks which minimizes the Earth-Mover's distance works the simplest in terms of improving the speech quality.

III. PROPOSED SYSTEM

The proposed system executes commands given by the user. Thus, it highly depends on just the voice commands given by the user to complete his job. Designing a voice-controlled computer interface not only makes the execution of a command easier, but it also helps the disabled individuals to control a computer. Hand-free computing is possible where a user can interact with the computer without the use of their hands. Speech recognition can be trained to recognise various voice commands. Disabled persons may find the hand-free computer is vital in their life. This system is designed to recognise the speech and execute with full capability Synthesising means it converts text to speech. The user is asked to provide voice command by using a microphone. The microphone shall take the speech as a command, and the analogue signals are converted to digital in the internal circuits. Then digital signals are processed as an acoustic model. Once the particular applications are identified by the system, it then opens the application. When the application is found, the system shall prompt the user to create a new application in the current working directory. The system would ask for various operation such as edit, read, paste, copy, and other similar operations that shall be performed inside it once the application opens. The system uses the ferment synthesis, a type of speech synthesiser for responding to the user through voice command.

IV. SYSTEM ARCHITECTURE

The speech recognition system accepts specific instructions once adequately trained. Sans usage of hands instructions is fed to the systems once the correctness is confirmed. They can be used by an inspector or engineer in a factory environment or even while driving.

A. Login

The authentication details are stored and the innards of the login module such as username, location, Gmail ID and password in a file; for instance, a notepad. In the login module, the trainer is allowed to produce trainer information which is used in the authentication process and also in providing necessary information to the other modules.

B. Synthesizer

A speech synthesiser converts transcription into speech. The synthesiser simplifies the process by figuring out a paragraph, sentence and any other structures of the sentence from the start till the end from the input text. Formatting data, punctuations are employed for several languages throughout this phase, and it also examines the input script for a particular paradigm of the language. Numerous mandated and unique actions necessitate for dates, times, numbers, currency amounts, email addresses, abbreviations, acronyms and lots of other forms in the English language. Thus, the demand to convert each expression to phonemes. A phoneme is a basic unit of sound in an exceeding language. American English has about 45 phonemes, together with the consonant and vowel sounds. Lastly, each sentence produces audio waveform adapting the phonemes and prosody information.

C. Affix commands

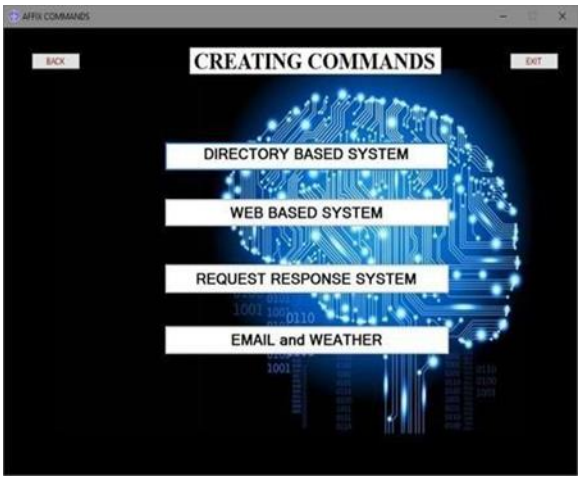
The three sorts of commands utilised in the system are Shell command, social command, and Web command. The storage of specific applications file and folder locations to the trainer is taken care of by Shell command. Colloquial languages are difficult to acknowledge for the Speech recogniser; thus, it demands to produce relevant commands. Any application incorporation performed is through with the assistance of this module. Employing web commands with the assistance of this module makes it simpler to access a person's default browser. The system's firewall and internet security conditions on how it has to be processed, for instance, a request-response system uses the social commands, which is active for "what" sort of questions.

D. Web Command

Uniform Resource Locator (URL) within the network is accessed using a web-based command system known as Web Command. Once added, any set of Uniform Resource Locator (URL) to the system, the net module provides with limited permissions to the system to be satisfied in order to access it. The login process needs to be done before accessing the net module. The trainer can only access the net module, and the user has consent to use the updated command by the trainer. The trainer is authorised to update commands while the user is generally an individual.

E. Shell Command

A directory-based system which deals with the directory of any file or action is known as Shell command. Thus, Shell command is one among significant augmentation made within the system for the following reason. The directory of the shell module could be a path to any variety of system and the only means to access it is by logging into the system, and which is possible only by the trainer. Thus, the trainer is the only one who can use this module to update the directory- based commands into the system with the help of a shell module.



V. WORKING OF MODEL

The work started with analyzing the audio commands given by the user through the microphone. This can be anything like getting any information, operating a computer's internal files, etc. This is an empirical qualitative study, based on reading above mentioned literature and testing their examples. Tests are made by programming according to books and online resources, with the explicit goal to find best practices and a more advanced understanding of Voice Assistant.

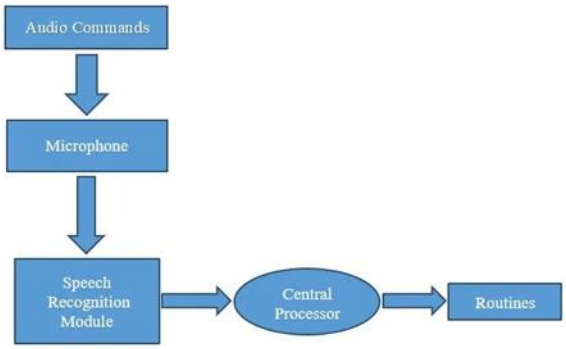
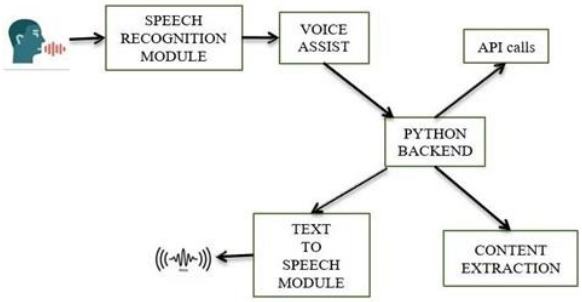


Fig. shows the workflow of the basic process of the voice assistant. Speech recognition is used to convert the speech input to text. This text is then fed to the central processor which determines the nature of the command and calls the relevant script for execution.

VI. METHODOLOGY OFVIRTUAL ASSISTANT USING PYTHON



- *Speech Recognition module*

The system uses Google's online speech recognition system for converting speech input to text. The speech input Users can obtain texts from the special corpora organized on the computer network server at the information centre from the microphone is temporarily stored in the system which is then sent to Google cloud for speech recognition. The equivalent text is then received and fed to the central processor.

- *Python Backend:*

The python backend gets the output from the speech recognition module and then identifies whether the command or the speech output is an API Call and Context Extraction. The output is then sent back to the python backend to give the required output to the user.

- *API calls*

API stands for Application Programming Interface. An API is a software intermediary that allows two applications to talk to each other. In other words, an API is a messenger that delivers your request to the provider that you're requesting it from and then delivers the response back to you.

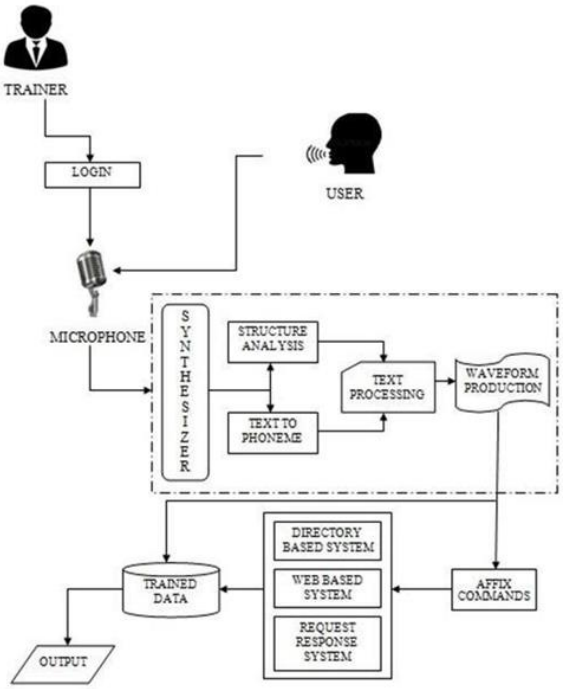
- *Content Extraction*

Context extraction (CE) is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents. In most cases, this activity concerns processing human language texts using natural language processing (NLP). Recent activities in multimedia document processing like automatic annotation and content extraction out of images/audio/video could be seen as context extraction TEST RESULTS.

- *Text-to-speech module*

Text-to-Speech (TTS) refers to the ability of computers to read text aloud. A TTS Engine converts written text to a phonemic representation, then converts the phonemic representation to waveforms that can be output as sound. TTS engines with different languages, dialects and specialized vocabularies are available through third-party publishers

CONCLUSION



In this paper "Virtual Assistant Using Python based on the concepts of AI-ML" we discussed the design and implementation of Digital Assistance.

The project is built using open-source software modules with PyCharm community backing which can accommodate any updates shortly. The modular nature of this project makes it more flexible and easier to add additional features without disturbing current system functionalities.

It not only works on human commands but also give responses to the user based on the query being asked or the words spoken by the user such as opening tasks and operations. It is greeting the user the way the user feels more comfortable and feels free to interact with the voice assistant. The application should also eliminate any kind of unnecessary manual work required in the user life of performing every task. The entire system works on the verbal input rather than the next one.