Intelligent Personal Assistant - Implementing Voice Commands enabling Speech Recognition

. Kumaran N[1], Rangaraj V[2],Siva Sharan S[3]and Dhanalakshmi R [4],
Department of Computer Science and Engineering
nkumaran811@gmail.com, ranagadon500@gmail.com,
sivasharan98@gmail.com, dhanalakshmisai@gmail.com

***Abstract*— intelligent personal assistant** (**IPA**) is a software agent performing tasks on behalf of an human or individual l based on commands or questions which are similar to chat bots. They are also referred as Intelligent Virtual Assistant which interprets human speech and respond via synthesized voices. IPAs and IVAs finds their usage in various applications such as home automation , manage to-do tasks and media playback through voice. **This paper aims to propose speech recognition systems and dealing with creating a virtual personal assistant. The existing system serves on the internet and is maintained by the third party. This application shall protect personal data from others and use the local database, speech recognition and synthesiser. A parser named SURR(Semantic Unification and Reference Resolution) is employed to recognise the speech. Synthesizer uses text to phoneme.**

Keywords-Intelligent Personal Assistant, Intelligent Virtual Assistant, Speech Recognition, Speech Synthesizer, Semantic Unification and Reference Resolution

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.

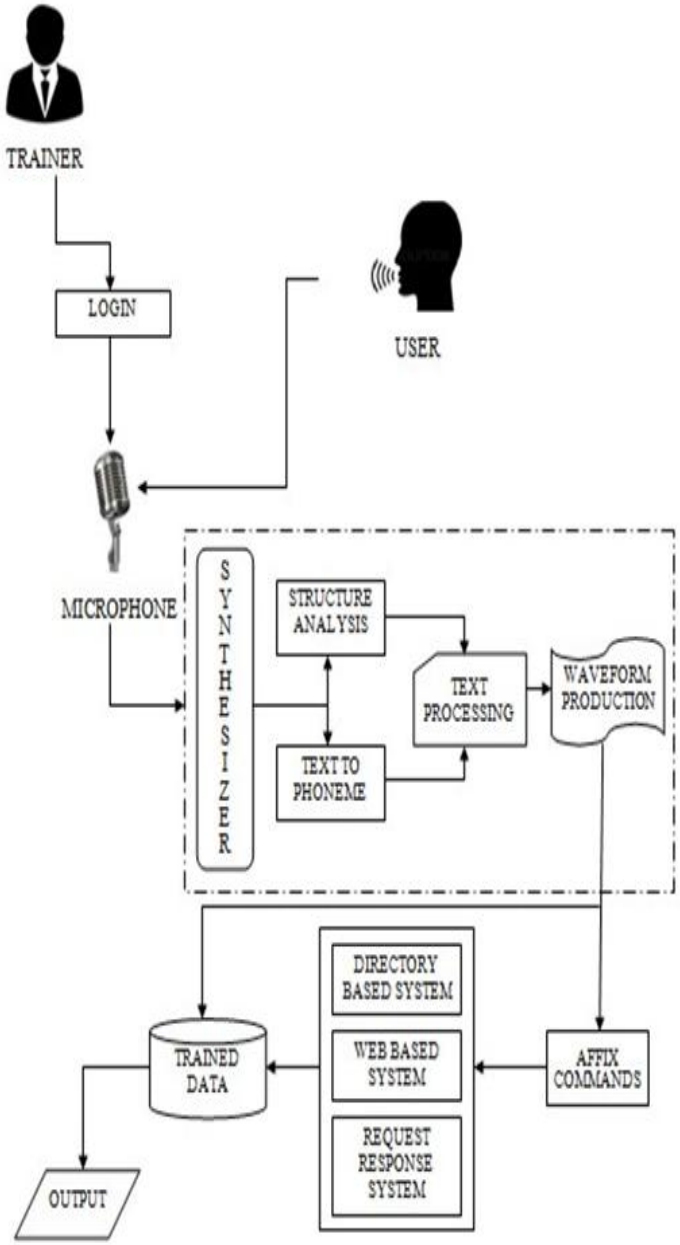
E. Shell Command

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.

F. Social Command

The request-response system which asks most of the "W" kind of question to the system is known as the social command as it forms an interactive system with the user. The updating of information in a social command is a continuous process. It is vast relating to the necessities of the user, as it might differ. The trainer frames questions for the user as it is attained from the requirements of the user by default. Furthermore, the trainer has the credibility to entree for updating the social command into the system for the reason that the login can be done only by the trainer. In contrast, the request-response system is more interactive module.



V. ALGORITHM

A. Parsing

This parser sets a label for the word utterance or parts of it as a semantic label.and the most commonly used techniques in parser are supported by CFG or probabilistic CFG which are made to support the speech data where it is collected and designed by developers. For our system, we implemented the Jurc˘ıcek ets al's algorithm. terns in chunks of the words in the sentence and the temporary SF (Semantic Frame) is inside it . The operation Semantic Frame is triggered after the pattern is detected. The foundation has two parts named trigger part and operation part which act as pattern detector and to apply transformation. It is automatically trained from the final Semantic Frame which is located in the annotated corpus. Every step in the process of training, the algorithm verifies the trigger-operation pairs in the parser. The pair which does not trigger any operation scores, i.e. the one whose application leads to the most critical Levenshtein distance reduction between the obtained SFs and hence the referred algorithm is added to decode the piles and applied to the corpus which is temporary.

B. Unification

As it was mentioned earlier, it is only the rules derived from a priori offline training along with the user's input that aid the parsing. Contextual information is not considered or processed for this and thus the results might be vague. For example, if the user enters "four" as an input, the system will identify it as "input: number = 4". Since the input itself doesn't inform us much, we require contextual information to support the semantic representation the system comes up with. This is why we have defined a Semantic Unification and Reference Resolution (SURR) module. The SURR generates and maintains a dynamic tree structure where nodes represent slots and the stems between nodes represent a transformation that applies to the slots. Thus, for a semantic frame to qualify, the system should be able to identify a path among these nodes. The system is said to have succeeded if it can find at least one path leading from the initial split to the foundation nodes.

The algorithm shown below illustrates the process, given that slots are subsequent sets within the input of the Semantic Frame. In the line 4, the splitting for Set is stored in memory which helps the algorithm to stop the testing of the identical partition twice and it comes out of the recursive function when all doesn't match. The SURR algorithm displays a "Solution is not found" alert when it's not able to get the path inside the structure. This can be used to alert the user about the utterances which are not understood by the system.

Algorithm 1 SURR algorithm

```
1: procedure FINDPATH(Slots)
2:     if Slots is made of root sets of slots then return Slots
3:     else
4:         split Slots into two parts, Split and Remaining
5:         if Split can be mapped to a node Node in the trees then
6:             if the transformation Transform from Node to its parent is valid then
7:                 apply Transform
8:                 merge the transformation of Split with Remaining to get a new set TransSlots
9:                 FINDPATH(Slots)
10:            else
11:                FINDPATH(Slots)
12:            end if
13:        else
14:            FINDPATH(Slots)
15:        end if
16:    end if
17: end procedure
```

Finally, the dialog act converter maps the contextualized meaning representation of the user input and the dialog acts. The reasoning engine maintains a stack of tasks and instructs them to retrieve expectations. After that, the system attempts to transform the frame into a parametrized act. If it fails to do so, it generates the "cannot map" act.

A. Text Pre-processing

The text is examined to identify particular language constructs. We require different kinds of treatments to construct terms for acronyms, abbreviations, time & date, currency, email IDs and other data formats.

B. Text to Phoneme

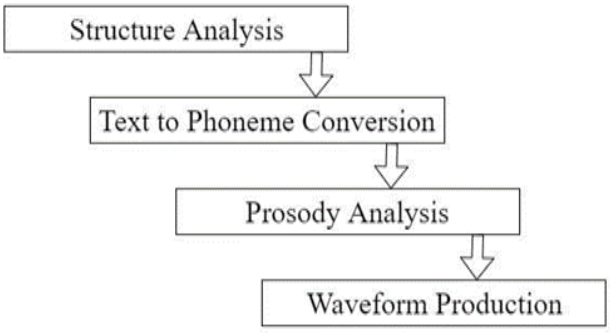
Each of the texts are transformed into phonemes. Simply speaking, a phoneme is a basic unit of sound in a language. For instance, the American English consists of around 46 phonemes (including the consonants and the vowels). To explain with an example, the word "Morning" comprises four phonemes. The number of phonemes, their styles and sounds vary according to the language. In the final step, the digital speech signal is converted into an analogue waveform.

C. Prosody Analysis

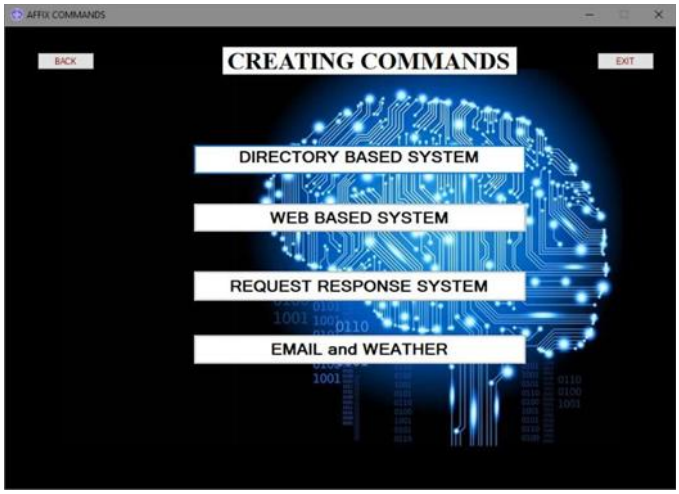
Prosody analysis is to verify the style and structure of the word in the sentence. phonemes make the proper sound appropriate to the sentence. it has various advantages over the sound of the word and speech being spoken. Prosody analysis includes the timing, the pausing, the pitch, the stress on words, the speaking rate, and other features.

D. Waveform Production

Ultimately, prosody and phonemes data are combined to produce the waveform of the text. Some ways are there to produce the waveform for the speech using the phoneme and prosody information.



V. EXPERIMENTAL RESULTS



VI. CONCLUSION

This system of Intelligent Personal Assistant – Virtual Digital Assistant using speech recognition and synthesizer finds its applications in several domains. The system is implemented with favorable results with the supported software development for speech recognition. Different tools and applications are integrated for the execution. Few other applications in the domain which can be integrated with the existing system can be online and automated learning. Robotic process Automation an also be implemented which requires speech recognition for few applications

REFERENCES

[1] F. Jurcˇ´ıcek, F. Mairesse, M. Gaˇ siˇ c, S. Keizer, B. Thom-´ son, K. Yu, and S. Young, "Transformation-based Learning for semantic parsing," in *Proceedings of INTERSPEECH*, 2009

[2] Francesco Nesta, Marco Matassoni, Ramon Fernandez Astudillo, "A Flexible Spatial Blind Source Extraction Framework For Robust Speech Recognition In Noisy Environments", Computer Speech and Language,

[3] H. Suzuki, H. Zen, Y. Nunkuku, C. Miyajima, K. Tokuda, and I. Kitumuru, "Speech Recognition Using Voice Characteristic Dependent Acoustic Models",Acoustics, Speech, and Signal Processing, 2003. Proceedings, May 2003.

[4] Jinmook Lee, Seongwook Park, Injoon Hong and HoiJun Yoo, "An Energy-efficient Speech Extraction Processor for Robust User Speech Recognition in Mobile Head-mounted Display Systems", IEEE Transactions on Circuits and Systems II, Volume: 64, Issue: 4, April 2017.

[5] Michael Price, *Member,* James Glass and Anantha P. Chandrakasan, "A Low-Power Speech Recognizer and Voice Activity Detector Using Deep Neural Networks", IEEE Journal of Solid-State Circuits, October 2017.

[6] Nguyen Duc Thang, Sungyoung Lee, Young-Koo Lee, Kyung Hee, "Fast Constrained Independent Component Analysis For Blind Speech Separation With Multiple References",Computer Sciences and Convergence Information Technology (ICCIT) Dec. 2010.

[7] Philipp Aichinger, Martin Hagmuller, Berit SchneiderStickler, Jean Schoentgen and Franz Pernkopf, "Tracking Of Multiple Fundamental Frequencies in Diplophonic Voices", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Volume: 26, Issue: 2, October 2017.

[8] Suma Swamy and K.V Ramakrishnan, "An Efficient Speech Recognition System", Computer Science & Engineering: An International Journal (CSEIJ), Vol. 3, No. 4, August 2013.

[9] Wei-Yeh Shih, Jui-Chieh Liao, Kuan-Ju Huang, Wai-Chi Fang, Gert Cauwenberghs, and Tzyy-Ping Jung, "An Efficient VLSI Implementation Of On-Line Recursive ICA Processor For Real-Time Multi-Channel EEG Signal Separation", Engineering in Medicine and Biology Society (EMBC), September 2013.

[10] Wei-Yeh Shih, Kuan-Ju Huang, Chiu-Kuo Chen, WaiChi Fang, Gert Cauwenberghs and Tzyy-Ping Jung, "An Effective Chip Implementation of A Real-time Eightchannel EEG Signal Processor Based on On-line Recursive ICA Algorithm",Biomedical Circuits and Systems Conference (BioCAS), 2012 IEEE, January 2013.

[11] Yuki Saito, Shinnosuke Takamichi and Hiroshi Saruwatari, "Statistical Parametric Speech Synthesis Incorporating Generative Adversarial networks",IEEE/ACM Transactions onAudio, Speech, andLanguage Processing, Volume: 26, Issue: 1, October 2017.