VOICE COMMAND EXECUTION WITH SPEECH RECOGNITION AND SYNTHESIZER

R Ram kumar[1], A Vimal kumar[2], R Deepa[3], S Shalini[4]

[1] & [2] UG scholar, [3] & [4]Assistant professor
Department of Computer Science and Engineering
Prince Dr.K.Vasudevan College of Engineering and Technology, Chennai, India.

Abstract - *Speech recognition is known as "Automatic Speech Recognition" (ASR), or "Speech To Text" (STT). To understand the speech recognition and its fundamentals with its working procedures. The applications of speech recognition in various areas are explored. The speech recognition system is implemented as a desktop application in the given system. The speech recognition system development can be mainly used for speech recognition, speech generation, text editing, and tool for operating machine through voice. The speech recognition system is also called as Automatic speech recognition. The text-to-speech is implemented in the given system. In English, special treatment is required for abbreviations, acronyms, dates, times, numbers, currency amounts, email addresses and many other forms. Other languages need special processing for these forms and most languages have other specialized requirements. The speech recognition must be dynamic, so that the execution time will be reduced. The synthesizer based algorithm is used to preprocess the input voice and generate relevant output.*

***Key Words*: Automatic Speech Recognition (ASR), Speech To Text (STT), speech generation, synthesizer.**

1. INTRODUCTION

Artificial Intelligence Technique is a manner to organize and use the knowledge efficiently in such a way that, it should be perceivable by the people who provide it. It should be easily modifiable to correct errors. It should be useful in many situations though it is incomplete or inaccurate. AI techniques elevate the speed of execution of the complex program it is equipped with. AI has been dominant in various fields such as Gaming, Natural Language Processing, Expert Systems, Vision Systems, Speech Recognition, Handwriting Recognition, Intelligent Robots etc. The speech recognition is the main category used under artificial intelligence. Computers are known to execute accurately every command given to it by the user. There are various commands that can be inputted to the computer. The command could be in various formats. For example the command could be to print a document or to play audio/video files or to open a file or to paint a picture etc.

Identifying the capability of speech recognition system and working to build its efficiency further to execute high level voice commands given by the user. This finds great advantage in the moving world where the user has to give only voice commands to finish his job. Consequently this application is expected to reduce the time delay in executing commands with GUI. Synthesizer algorithm is used in the proposed system. It is Hands-free computing. Reducing the time delay when compared to the existing speech recognition system. User has to give only voice commands to finish his task. Three types of commands such as social commands, web commands and shell commands are used in the proposed system. These commands can only be updated using the trainer whereas user can only use the commands. The trainer is authenticated using a security level to access the updating of commands into the system.

2. RELATED WORK

[2] This paper describes digital circuit architectures for Automatic Speech Recognition (ASR) and Voice Activity Detection (VAD) with improved accuracy, programmability, and scalability. Our ASR architecture is designed to minimize off-chip memory bandwidth, which is the main driver of system power consumption. A SIMD processor with 32 parallel execution unit's efficiently evaluates feed-forward Deep Neural Networks (DNNs) for ASR, limiting memory usage with a sparse quantized weight matrix format.

[3] Diplophonia is a type of pathological voice in which two fundamental frequencies are present simultaneously. Specialized audio analyzers that can handle up to two f0 s in diplophonic voices are in their infancy. We propose the tracking of up to two fs in diplophonic voices by Audio Waveform Modeling (AWM), which involves obtaining candidates by repetitive execution of the Viterbi algorithm, followed by waveform Fourier synthesis, and heuristic candidate selection with majority voting. The fast variant is more than twice as fast as the fastest relevant benchmark, and the median error rate is 9.52%.Furthermore, illustrative results of connected speech analysis are reported.

[4] A method for statistical parametric speech synthesis incorporating Generative Adversarial Networks (GANs) is proposed. Although powerful Deep Neural Networks (DNNs) technique can be applied to artificially synthesize speech waveform, the synthetic speech quality is low compared with that of natural speech. A GAN introduced

in this paper consists of two neural networks: a discriminator to distinguish natural and generated samples, and a generator to deceive the discriminator. In the proposed framework incorporating the GANs, the discriminator is trained to distinguish natural and generated speech parameters, while the acoustic models are trained to minimize the weighted sum of the conventional minimum generation loss and an adversarial loss for deceiving the discriminator. We investigated the effect of the divergence of various GANs, and found that a Wasserstein GAN minimizing the Earth-Mover's distance works the best in terms of improving synthetic speech quality.

[5] Blind Source Extraction (BSE) is an attractive approach to enhance multichannel noisy speech data, as a preprocessing step for an automatic speech recognition system. In this work were viewed the BSE architecture and improved each system block in the framework in order to increase its flexibility and degree of blindness. To improve the overall performance, the output of the enhancement algorithm is then combined with robust ASR systems based on gammatone features analysis and on uncertainty decoding. Results obtained with different front-end and back-end configurations demonstrate the advantages of the proposed approaches.

[6] This paper involves representation an efficient VLSI implementation of on-line recursive ICA (ORICA) processor for real-time multi-channel EEG signal separation. The proposed design contains a system control unit, a whitening unit, a singular value decomposition unit, a floating matrix multiply and, and an ORICA weight training unit. The design of the ORICA processor is a mixed architecture, which is designed as different hardware parallelism according to the complexity of processing units. The shared arithmetic processing unit and shared register can reduce hardware complexity and power consumption. The proposed design is implemented used TSMC0nm CMOS technology with 8-channel EEG processing in 128Hz sample rate of raw data and consumes 2.827 mW at 50 MHz lock rate. . The realtime Multi-channel EEG signal separation has the performance of the proposed design is also shown to reach 0.0078125 s latency after each EEG sample time, and he average correlation coefficient between the original source signals and extracted ORICA signals for each 1s frame is 0.9763.

3. EXISTING SYSTEM

The speech recognizer is found in some operating system as a secondary option. But it is not used upto its full capacity and capability. The application of this inbuilt speech recognizer is limited but it has many scopes. This operates with 3 steps. The user has to setup the recognizer. After the successful installation and setup of the speech recognizer the speech recognition engine is displayed on the desktop. To use the recognizer a command "show numbers" is to be given. Then many numbers at random will be shown by default. The user has to then select the number to perform the desired operation. This obviously increases the time of execution is the drawback of this system.

Early speech recognition systems tried to apply a set of grammatical and syntactical rules to speech. If the words spoken fit into a certain set of rules, the program could determine what the words were. However, human language has numerous exceptions to its own rules, even when it's spoken consistently. Accents, dialects and mannerisms can vastly change the way certain words or phrases are spoken. Imagine someone from Boston saying the word "barn." He wouldn't pronounce the "r" at all, and the word comes out rhyming with "John." Or consider the sentence, "I'm going to see the ocean." Most people don't pronunciate their words very carefully. The existing system deals with third party application with an secondary view i.e. the speech recognition system is used as secondary, whereas the external applications can be accessed only using the command "show numbers" where the action to be done is stored.

4. PROPOSED SYSTEM

Identifying the capability of speech recognition system and working to build its efficiency further to execute high level voice commands given by the user. This finds great advantage in the moving world where the user has to give only voice commands to finish his job. Consequently this application is expected to reduce the time delay in executing commands with GUI. Synthesizer algorithm is used in the proposed system. It is Hands-free computing. Reducing the time delay when compared to the existing speech recognition system. User has to give only voice commands to finish his task. Three types of commands such as social commands, web commands and shell commands are used in the proposed system. These commands can only be updated using the trainer whereas user can only use the commands. A speech synthesizer converts written text into spoken language. Speech synthesis is also referred to as Text-To-Speech (TTS) conversion. Process the input text to determine where paragraphs, sentences and other structures start and end. For most languages, punctuation and formatting data are used in this stage. Analyze the input text for special constructs of the language. In English, special treatment is required for abbreviations, acronyms, dates, times, numbers, currency amounts, email addresses and many other forms. Convert each word to phonemes. A phoneme is a basic unit of sound in a language. US English has around 45 phonemes including the consonant and vowel sounds. Finally, the phonemes and prosody information are used to produce the audio waveform for each sentence. There are many ways in which the speech can be produced from the phoneme and prosody information.

A voice-controlled human intelligence based computer interface has been designed in a way that enables handicapped individuals to operate a computer. Hands-free computing is a user interface in which user can work without the use of hands, a common requirement of human interface devices such as the mouse and keyboard are made secondary in the proposed system. Speech recognition system can be trained to recognize specific commands and upon confirmation of correctness, instructions can be given to systems without the use of hands. This may be useful while driving or to an inspector or engineer in a factory environment. The proposed system can answer complex questions or task given by the user as it performs the action in quicker time delay. Energy is saved efficiently as the performance of the system is increased.

System Architecture

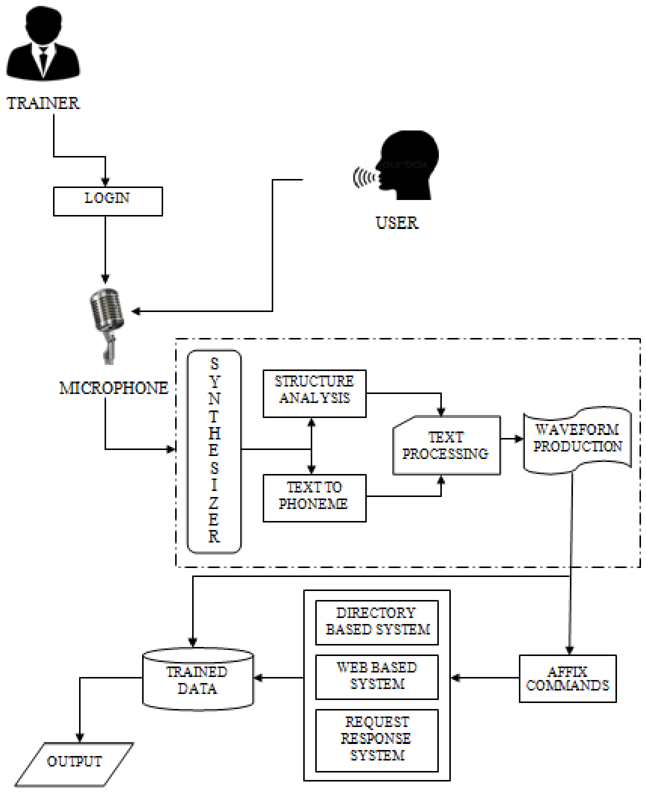


Fig -1: System Architecture

Design Engineering deals with the various diagrams for the implementation of project. Design is the creation of a plan or convention for the construction of an object, system or measurable human interaction such as the example terms given as follows in architectural based design blueprints, engineering drawings, business processes, circuit diagrams, and sewing based patterns sets. Design has different connotations in different fields. Software design is a process through which the requirements are translated into representation of the model.

This software is designed to recognize the speech and also has the capabilities for speaking and synthesizing means it can convert speech to text and text to speech. The user is asked to provide voice command via the microphone. The microphone intakes the command and the analog signals are converted to digital ones in the internal circuit. These digitized signals are processed as acoustic model. The program is essentially executed at run-time as the given program is dynamic.

Login

The module allows the trainer to provide the trainer information for authentication. It provides all basic information for other modules. The authentication details are stored in text file (i.e.) notepad. The contents of the login module are location, username, Gmail ID and password. The user given credentials are stored onto the file embedded to the code. The authentication is only done for the trainer.

Synthesizer

A speech synthesizer converts written text into spoken language. Speech synthesis is also referred to as Text-To-Speech (TTS) conversion. Process the input text to determine where paragraphs, sentences and other structures start and end. For most languages, punctuation and formatting data are used in this stage. Analyze the input text for special constructs of the language. In English, special treatment is required for abbreviations, acronyms, dates, times, numbers, currency amounts, email addresses and many other forms. Convert each word to phonemes. A phoneme is a basic unit of sound in a language. US English has around 45 phonemes including the consonant and vowel sounds. Finally, the phonemes and prosody information are used to produce the audio waveform for each sentence. There are many ways in which the speech can be produced from the phoneme and prosody information.

Affix commands

There are three types of commands used in the system. They are Shell command, Social command, and Web command. The shell command stores Location of every file, folder and application is to be specified initially to trainer. It is recommended to provide related commands. Colloquial languages are difficult to recognize for speech recognizer. Any application integration can be done with the help of this module. Using web command, web pages can be accessed using your default web browser with the help of this module. It processes as per the system's firewall and internet security. The social command is used for request-response system, which is used for "what" type of questions.

Directory based system

Shell command is the directory based system. Shell command is one of major enhancement made in the system. The shell command deals with the directory of any particular file or action. The directory is a path or way to access any type of file from the system. The process of accessing the shell command module is by logging in the system. As the trainer can only login into the system the trainer can only use this module. The trainer can only update the directory based commands into the system with the use of shell command module.

Web based system

The web command is the web based command system. It is enhanced to mainly access the Uniform Resource Locator (URL) in the network. Any set of Uniform Resource Locator (URL) can be added to the system. The web command module has some permission to be satisfied to access it. The login must be done before accessing the web command module. As the trainer can only login the system, the web command module can only be accessed by the trainer. The user has the permission to use the updated command by the trainer. Mostly user is considered as handicapped person, hence trainer is given the access to update commands.

Request response system

Social command is the request response system, which is the "W" type of question asked to the system. An interactive system with the user is made by the request response system. Social command is the vast type of command whereas updation of data is a continuous process, as the requirements of the user may vary. The questions framed for the user as default are obtained from the user given requirements to the trainer. The trainer has the credibility to access the updation of social command into the system, as the login can only be done by the trainer. The request response system is more interactive module.

5. ALGORITHM

TEXT PRE-PROCESSING

Analyze the input text for special constructs of the language. In English, special treatment is required for abbreviations, acronyms, dates, times, numbers, currency amounts, email addresses and many other forms. The remaining steps convert the spoken text to speech.

TEXT-TO-PHONEME CONVERSION

Convert each word to phonemes. A phoneme is a basic unit of sound in a language. US English has around 45 phonemes including the consonant and vowel sounds. For example, "times" is spoken as four phonemes "t ay m s". Different languages have different sets of sounds (different phonemes). The remaining steps convert the digitalized voice signal to the analog waveform.

PROSODY ANALYSIS

Process the sentence structure, words and phonemes to determine appropriate prosody for the sentence. Prosody includes many of the features of speech other than the sounds of the words being spoken. This includes the pitch (or melody), the timing (or rhythm), the pausing, the speaking rate, the emphasis on words and many other features.

WAVEFORM PRODUCTION

Finally, the phonemes and prosody information are used to produce the audio waveform for each sentence. There are many ways in which the speech can be produced from the phoneme and prosody information.

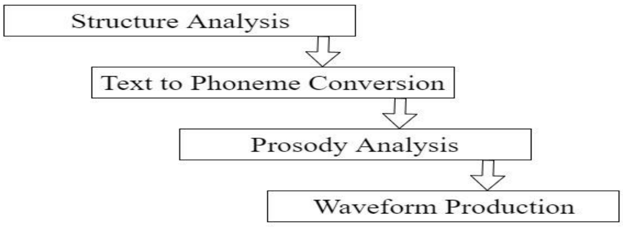
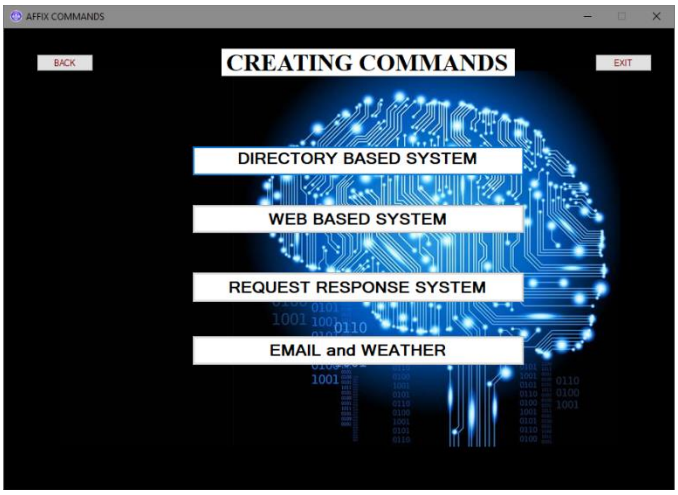


Fig -2: Synthesizer Structure

6. EXPERIMENTAL RESULTS

Microsoft .NET is a set of Microsoft software technologies for rapidly building and integrating XML Web services, Microsoft Windows-based applications, and Web solutions. The .NET Framework is a language-neutral platform for writing programs that can easily and securely interoperate. There's no language barrier with .NET: there are numerous languages available to the developer including Managed C++, C#, Visual Basic and Java Script. ADO.NET is based on the fundamental architecture of .NET Framework. The .NET Framework is an integral Windows component that supports building and running the next generation of applications and XML Web services. The .Net Framework includes mainly three Data Providers for ADO.NET. To provide a code-execution environment that minimizes software deployment and versioning conflicts. SQL Server uses the SQL Connection object, OLEDB uses the OLEDB Connection Object and ODBC uses ODBC Connection Object respectively.



7. CONCLUSION AND FUTURE ENHANCEMENT

The Proposed System of speech recognition started with a brief introduction of the technology and its applications in different sectors. The project part of the Report was based on software development for speech recognition. At the later stage we discussed different tools for bringing that idea into practical work. After the development of the software finally it was tested and results were discussed, few deficiencies factors were brought in front. The future enhancement involves automatic learning. After the testing work, advantages of the software were described and suggestions for further enhancement and improvement were discussed.

8. REFERENCES

[1] Jinmook Lee, Seongwook Park, Injoon Hong and Hoi-Jun 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.

[2] 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, Volume: 53, Issue: 1, October 2017.

[3] Philipp Aichinger, Martin Hagmuller, Berit Schneider-Stickler, 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.

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

[5] 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, Volume: 27, Issue: 3, May 2013.

[6] 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.

[7] 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.

[8] 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.

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

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