HANDLING SYSTEM APPLICATIONS THROUGH SPEECH RECOGNITION

Praveen A L[1], Gireesh Babu C N[2]

[1]*Department of Information Science Engineering, B M S I T, Bangalore (India)*

[2]*Assistant Professor, Department of Information Science Engineering, B M S I T, Bangalore (India)*

ABSTRACT

The evolution and development of personal computers & laptop automation is progressing towards the future of creation of an ideal smart environment. Optionally, Voice/speech automation system design has been developed considering the busy schedules of an individual and also giving a special attention to people with disabilities, sick patients etc. Thus, providing a suitable control scheme using voice/speech communication mode can help them do their daily routines. Besides, voice control access would be used as command for a better purpose. In this, we like to tell a smart automation system using voice/speech recognition. The scope of this research work includes controlling, monitoring personal computers & laptop applications, browser applications, E-mailing & other features from Graphical User Interface (GUI) using Microsoft Visual Basic software that uses Speech Recognition engine as an input source. The research methodology involved in this application knowledge, is in the field of voice frequency communication and computer programming

Keywords*: *Automated Speech, HMM Module, Speech Automation, Speaker Dependent, Speaker Independent.

1. INTRODUCTION

A. Definition Of Speech Recognition

Speech Recognition (is also known as automatic speech Recognition (ASR), or computer speech recognition) is the process of converting a speech signal to a sequence of words, by means of an algorithm implemented as a computer program.

1.1 Speech Recognition

It is an ability of a computer, computer software program, or a hardware device to decode the human voice into digitized speech that can be interpreted by the computer or hardware device. Speech recognition is used to perform commands, mouse, write without having to operate a keyboard, operate a device, or use any buttons. This can be done on a computer using (ASR) Automatic Speech Recognition software programs. Many ASR programs require the user to "train" the ASR program to recognize their voice so that it can more accurately convert the speech to text. For example, a user could say "open browser" and the computer would open an

Internet browser and allow that user to browse the Internet. Fig 1 below shows the basic model of a Speech Recognition system [1].

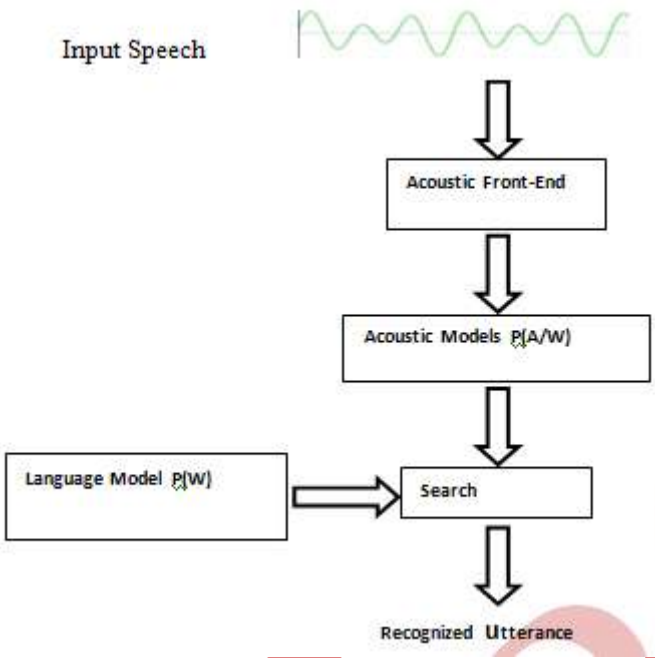

Fig 1: Basic model of Speech Recognition

1.2 Types Of Speech Recognition Systems

• **Speaker dependent**- However, generally require an extended training session during which the computer system becomes accustomed to a particular voice and accent. Such systems are said to be speaker dependent. A speaker dependent system is developed to operate for a single speaker [1].

• **Speaker independent** - A speaker independent system is developed to operate for any speaker of a particular type (e.g. American English). These systems are the most difficult to develop, most expensive and accuracy is lower than speaker dependent systems. However, they are more flexible. Speaker–independent software is designed to recognize anyone's voice, so no training is involved [1].

• **Speaker adaptive** - A third variation of speaker models is now emerging, called speaker adaptive. Speaker adaptive systems usually begin with a speaker independent model and adjust these models more closely to each individual during a brief training period.

II. LITERATURE SURVEY

Speech recognition is the ability of a machine or program to identify words and phrases in spoken language and convert them to a machine-readable format. Many speech recognition applications, such as voice dialing, simple data entry and speech-to-text are in existence today. Attempts to build automatic speech recognition (ASR) systems were first made in the 1950s. These early speech recognition systems tried to apply a set of grammatical and syntactical rules to identify speech. If the spoken words adhered to a certain rule set, the system could recognize the words [7].

The survey of various papers referenced is as below:

• S.K.Katti et al have worked on "Speech recognition by machine: A review". This paper presents a brief survey on ASR, provides a technological perspective and an appreciation of the fundamental progress that has been accomplished in the area of speech recognition [1]

• Heather Sobey et al has worked on "Literature Survey-Automatic Speech Recognition". In this paper, one of the problems faced in speech recognition that is the spoken word vastly altered by accents, dialects and mannerisms is addressed. [2]

• Preeti Saini et al have worked on "Automatic Speech Recognition: A Review". This paper presents a study of basic approaches to speech recognition and their results shows better accuracy [3].

• Yunxin Zhao et al in the paper on "A Speaker-Independent Continuous Speech Recognition System using Continuous Mixture Gaussian Density HMM of Phoneme-Sized Units". In this paper a bottom-up merging algorithm for generating mixture Gaussian density from models of sub-segments of phone units is developed, which we have incorporated in our work and this has proven successful in enhancing the recognition performance when comparing with the baseline of the segmental K-means.[5]

III. PRESENT WORK

In this paper, we proposed a modeled which can handle the desktop applications as well as browser functions through speech. In this paper, we proposed models to include some default commands to open and close a window, to know current date/time, to run audio/video files, to get the weather information [3]. We proposed the models to include the commands to open web pages. A special attention has given to the E-mail functions (Commands to compose and sending a mail, to read a mail, to open an inbox, drafts, sent mails list). The proposed application has a special feature that, a user's can add their own commands in their own way to understand easily. In this paper we studied how to trap human voice in a digital computer and decode it into corresponding text. Through this paper, we present a scheme to convert speech to text. The key factor in designing such system is the target audience. For example, physically handicapped people should be able to wear a headset and have their hands and eyes free in order to operate the system.

IV. BLOCK DIAGRAM

A block diagram of a complete Automatic speech recognition system is comprised of modules as shown in the Fig. 2 below. The first step in the processing is inputting the data or information through any kind of input signal in as a speech input. The second step in the recognizer is a combined word-level/sentence-level matching procedure. The way this is accomplished is as follows. Using a set of sub word models along with a word lexicon, a set of word models is created by concatenating each of the sub word models as specified by the word lexicon. The word-level match procedure provides scores for individual words as specified by the sentence-level match procedure and the semantics.

4.2.1 Speech Signal acquisition

At this stage, Analog speech signal is acquired through a high quality, noiseless, unidirectional microphone in wav format and converted to digital speech signal [4].

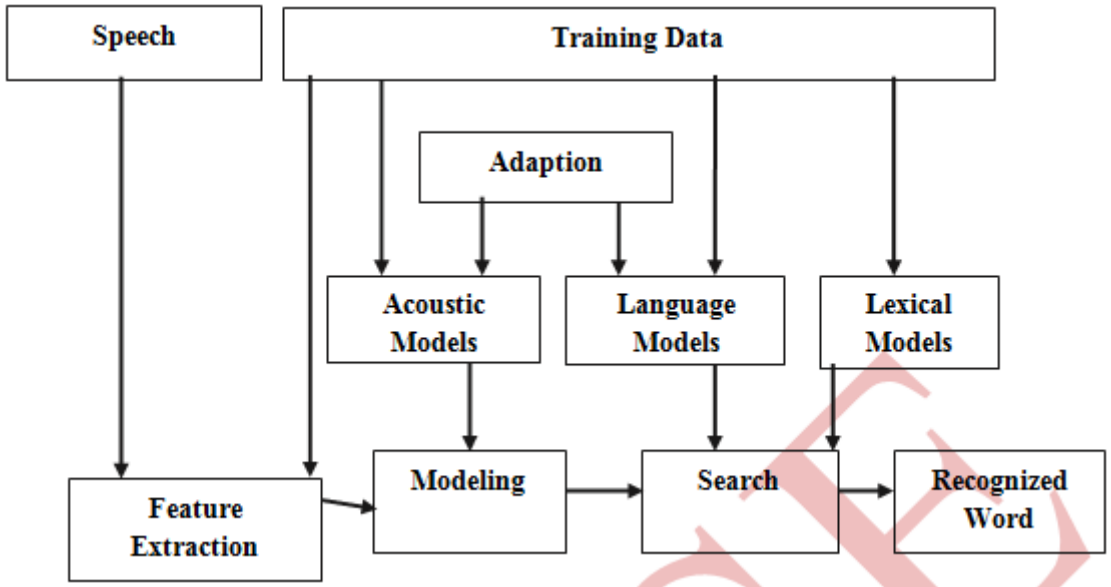

Fig 2: Block Diagram of Speech Recognition

4.2.2 Feature Extraction

Feature extraction is a very important phase of ASR development during which a parsimonious sequence of feature vectors is computed so as to provide a compact representation of the given input signal. Speech analysis of the speech signal acts as first stage of Feature extraction process where raw features describing the envelope of power spectrum are generated. An extended feature vector composed of static and dynamic features is compiled in the second stage. Finally this feature vector is transformed into more compact and robust vector. Feature extraction, using MFCC, is the famous technique used for feature extraction [4].

4.2.3 Acoustic Modeling

Acoustic models are developed to link the observed features of the speech signals with the expected phonetics of the hypothesis word/sentence. For generating mapping between the basic speech units such as phones, tri-phones & syllables, a rigorous training is carried. During training, a pattern representative for the features of a class using one or more patterns corresponding to speech sounds of the same class.

4.2.4 Language & Lexical Modeling

Word ambiguity is an aspect which has to be handled carefully and acoustic model alone can't handle it. For continuous speech, word boundaries are major issue. Language model is used to resolve both these issues. Generally ASR systems use the stochastic language models. These probabilities are to be trained from a corpus. Language accepts the various competitive hypotheses of words from the acoustic models and thereby generates a probability for each sequence of words. Lexical model provides the pronunciation of the words in the specified language and contains the mapping between words and phones. Generally a canonical pronunciation available in ordinary dictionaries is used. To handle the issue of variability, multiple pronunciation variants for each word are covered in the lexicon but with care. A G2P system- Grapheme to Phoneme system is applied to better the

performance the ASR system b predicting the pronunciation of words which are not found in the training data [4].

4.2.5 Model Adaptation

The purpose of performing adaptation is to minimize the system's performance dependence on speaker's voice, microphones, transmission channel and acoustic environment so that the generalization capability of the system can be enhanced.

4.2.6 Recognition

Recognition is a process where an unknown test pattern is compared with each sound class reference pattern and, thereby, a measure of similarity is computed.

4.3 Working

As the Web transforms from a text only medium into a more multimedia rich medium the need arises to perform searches based on the multimedia content. In this paper, we present an audio search engine to tackle this problem. The engine uses speech recognition technology to index spoken audio, when no transcriptions are available. If transcriptions (even imperfect ones) are available we can also take advantage of them to improve the indexing process.

Our engine indexes several thousand talks and news radio shows covering a wide range of topics and speaking styles from a selection of public web sites with multimedia archives [4].

4.3.1 Methodology

Speech Recognition is technology that can translate spoken words into text. Some SR systems use "training" where an individual speaker reads sections of text into the SR system. These systems analyze the person's specific voice and use it to fine tune the recognition of that person's speech, resulting in more accurate transcription. Systems that do not use training are called "Speaker Independent" systems. Systems that use training are called "Speaker Dependent" systems. The term voice recognition refers to finding the identity of "who" is speaking, rather than what they are saying. Recognizing the speaker voice recognition can simplify the task of translating speech in systems that have been trained on specific person's voices or it can be used to authenticate or verify the identity of a speaker as part of a security process. "Voice recognition" means "recognizing by voice", something humans do all the time over the phone. As soon as someone familiar says "hello" the listener can identify them by the sound of their voice alone [5]. Speech synthesis is the artificial production of human speech. A computer system used for this purpose is called a speech synthesizer, and can be implemented in software or hardware. A text-to-speech (TTS) system converts normal language text into speech; other systems render symbolic linguistic representations like phonetic transcriptions into speech. These features of the speech recognition and synthesis combined together to form a library for perform both recognition and synthesis process in a single machine. The user can search the content by giving the work likewise the human can give the keyword through the speech sequence the system can analysis and understand the human need and search the content and parse the raw content to extract the exact knowledge data [5].

4.3.1.1 Data Preparation

The first stage of any recognizer development project is data preparation. Speech data is needed both for training and for testing. In the system built here, all of this speech was recorded from scratch. The training data is used during the development of the system. Test data provides the reference transcriptions against which the recognizer's performance can be measured and a convenient way to create them is to use the task grammar as a random generator.

4.4 Modular Description

4.4.1 Speech Synthesis Module

A speech synthesizer takes text as input and produces an audio stream as output. Speech synthesis is also referred to as text-to-speech (TTS).

A synthesizer must perform substantial analysis and processing to accurately convert a string of characters into an audio stream that sounds just as the words would be spoken. The easiest way to imagine how this works is to picture the front end and back end of a two-part system.

4.4.1.1 Text Analysis

The front end specializes in the analysis of text using natural language rules. It analyzes a string of characters to determine where the words are (which is easy to do in English, but not as easy in languages such as Chinese and Japanese). This front end also figures out grammatical details like functions and parts of speech. For instance, which words are proper nouns, numbers, and so forth; where sentences begin and end; whether a phrase is a question or a statement; and whether a statement is past, present, or future tense.

4.4.1.2 Sound Generation

The back end has quite a different task. It takes the analysis done by the front end and, through some non-trivial analysis of its own, generates the appropriate sounds for the input text. Older synthesizers (and today's synthesizers with the smallest footprints) generate the individual sounds algorithmically, resulting in a very robotic sound. Modern synthesizers, such as the one in Windows Vista and Windows 7, use a database of sound segments built from hours and hours of recorded speech [11].

The Following steps summarizes how the Speech Synthesis works:-

- The text in any file or a source can then converted to the voice through synthesizer.
- It will convert the digital string to the respective voice through the US or UK Dictionary which we were used.
- Then it will dictate the related word in the dictionary.

4.4.2 Speech Recognition Module

- Speech recognition is the process of converting spoken language to written text or some similar form.
- To get the command or voice from the user through microphone and recognize the acoustics.

- Then it will check the clarity of voice and pass to the Extraction part.
- We use the globalized grammar database using "Google API" via internet to perform Speech Recognition.
- Speech recognition is the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words.
- The recognized words can be the final results for linguistic processing in order to reply the persons.

The major steps of a typical speech recognizer are:

1. Grammar
2. Signal processing
3. Phoneme recognition
4. Word recognition.

4.4.3 Data Extraction and Knowledge Understand Module

The data extraction is been done in the Speech Recognition system and how the understanding of different module is being done.

The voice signal can be then converted to the digital format based on the Parser.

- It will convert the digital string to word.
- Then it will process the command given by the user and sent to the synthesizer.
- It redirects to synthesizer.
- Data Extraction performs the role of getting the raw factor of the user input from the internet. These raw factor (XML Based data) could be mined and get the exact knowledge. Knowledge Understand can understand the user keyword to perform both searching process as well as command processing. Knowledge Understand module will understand the user's keyword to perform both searching process as well as command processing methods [7].

4.4.4 Command Processing

After understanding the knowledge, it does the correct task, what the user needs. Then the interaction processes simultaneously speak with user for effective communication and for verifying the tasks.

- After understand the knowledge it can do the correct task, what the user need.
- Then the interaction processes simultaneously speak with user for effective communication and verify the tasks.

All of these elements are critical to the selection of appropriate thread for words, phrases, and sentences. Consider that in English, a question usually ends with a rising pitch, or that the word "read" is pronounced very differently depending on its tense. Clearly, understanding how a word or phrase is being used is a critical aspect of interpreting text into sound. To further complicate matters, the rules are slightly different for each language. So, as you can imagine, the front end must do some very sophisticated analysis [7].

V. CONCLUSION

We presented a model that can handle all the system applications, web applications. A special attention is given to the E-mailing, User can compose, send and read the mails through voice only. User can able to write the text through both keyboard and voice input. Voice recognition of different default commands such as open an application, save and close a document have been included in this model. We can open different windows software's, based on voice input. User can play audio/video files through voice only. We propose a model to get the weather, temperature information and also know current date and time through voice only. A Windows Speech Recognition (WSR) tool is added to this model, to recognize a speaker voice. SAPI tool is included in this model to convert from speech to text. This model has a special feature that, a user's can add their own commands in their own way to understand easily

It is capable of meeting the requirements of people who are not able to use their hands. Our intention is to help the disabled people to get the benefits of advances in computer and electronics technologies. The primary aim of this model was to produce a basic system that would use a speech input for operating a computer. Requires less consumption of time in writing text.

REFERENCES

[1] M.A. Anusuya, S.K. Katti, "Speech Recognition by Machine: A Review", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 6, No. 3, 2009.

[2] Heather Sobey, "Literature Survey - Automatic Speech Recognition".

[3] Preeti Saini, Parneet Kaur, "Automatic Speech Recognition: A Review", International Journal of Engineering Trends and Technology- Volume4 Issue2- 2013.

[4] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, and Andrew W. Senior, "Recent Advances in the Automatic Recognition of Audio-Visual Speech".

[5] Yunxin Zhao, "A Speaker-Independent Continuous Speech Recognition System Using Continuous Mixture Gaussian Density HMM of Phoneme-Sized Units", IEEE Transactions On Speech And Audio Processing, Vol. 1, No. 3, July 1993.

[6] Sadaoki Furui, "50 years of Progress in speech and Speaker Recognition Research" , ECTI Transactions on Computer and Information Technology,Vol.1. No.2 November 2005.

[7] Shahramkhadivi and Hermann Ney, "Integrating of Speech recognition and machine translation in computer Assisted Translation", IEEE Transactions On Audio,Speech And Language Processing, Vol.16,No.1 Jan 2008.

[8] Rajesh Kumar Aggarwal and M. Dave, "Acoustic modeling problem for automatic speech recognition system: advances and refinements Part (Part II)", Int J Speech Technol, pp. 309– 320, 2011.

[9] Anusuya, M. A., & Katti, S. K. "Front end analysis of speech recognition: A review", International Journal of Speech Technology,Springer, vol.14, pp. 99–145, 2011.

[10] Foad Hamidi, Melanie Baljko, "Automatic Speech Recognition: A Shifted Role in Early Speech Intervention?"

[11] Shipra J, Hisar Rishi Pal Singh, "Automatic Speech Recognition: A Review", International Journal of Computer Applications (0975 – 8887) Volume 60– No.9, December 2012.

[12] Parwinder pal Singh, Er. Bhupinder singh, "Speech Recognition as Emerging Revolutionary Technology", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 10, October 2012.

Biographical Notes

Mr. Praveen A. L is presently pursuing B E. final year in Information Science Engineering Department from B M S Institute of Technology, Bangalore, Karnataka, India.

Mr. Gireesh Babu C.N is working as an Assistant Professor in Information Science Engineering Department from B M S Institute of Technology, Bangalore, Karnataka, India.