Mouse Movement using Speech and Non-Speech Characteristics of Human Voice

Sandeep Kaur

Abstract— In the modern era, mouse control has become an important part of human computer interaction which is difficult for physically disabled people. This research paper presents a system called as Vocal Mouse (VM). This device will allow users to continuously control the mouse pointer using words as well as sounds by varying vocal parameters such as vowel quality, loudness and pitch. Traditional method of using only standard spoken words was inefficient for performing continuous tasks and they are often recognized poorly by automatic speech recognizers. Now, VM will allow users to work on both continuous and discrete motion control. This includes commands given as words or regular sounds consisting of vowels and consonants. Low-level acoustic features are extracted in real time using LPC (Linear Predictive Coding). Pattern recognition is performed using a new proposed technique called "minimum feature distance technique". This proposed technique is based on calculating distances between the spoken word and each stored word in the library during training process. Features from pattern recognition module are processed to produce output in the form of cursor's 2-D movement. VM can be used by novice users without extensive training and it presents a viable alternative to existing speech-based cursor control methods.

Index Terms— acoustic features, continuous speech recognition, minimum feature distance, motor impairment, vocal sounds.

I. INTRODUCTION

Despite a lot of advancement in speech recognition technology over many years, the human voice still remains largely unexploited. Voice input has a number of potential benefits, especially for physically disabled people, one of the major limitations of current speech-based interaction methods is their inability to provide fluid and continuous input. This research paper will describe a new approach to implement the Vocal Mouse project by using both speech and non-speech characteristics of human voice to enable fluid hands-free computing.

It is important to enhance the expressivity of voice-based interaction due to number of reasons. People who are suffering from physical disabilities can't use standard input devices like keyboard and mouse, so they have only one option for gaining access to the computer i.e. hands-free input methods. In the United States alone, there are over 700,000 people with disabilities of the spinal cord, 70% of them are unemployed [1]. These people have limited mobility and motor control so they have few options available to access computers, to obtain or retain employment, to stay connected with people and gather information around them. In the nut shell, such individuals are not able to show their creativity. These issues extend to people with other motor impairments as well, including the 46 million adults in the United States diagnosed with arthritis, the one million with Parkinson's disease, and the 50,000 children and adults with muscular dystrophy [1].

Vocal Mouse as an enhanced voice-based interaction will also give benefit to people without motor impairments who find themselves in impairing situations. In situations like driving or interacting with a wall-sized display, hands-free interaction can be more suitable than traditional manual input devices. In desktop applications such as computer aided design tools that demand multiple dimensions and simultaneous channels of input, voice input can serve as an additional input modality to augment the standard keyboard and mouse interaction for greater control.

One of the major limitations of the current voice-based input recognition method is that they can give input only in the form of words as commands. User's vocal utterances are processed at the word level, which results in discrete interaction. This discrete motion could not be used for performing tasks like scrolling, zooming etc which requires input in the continuous form for continuous motion. This limitation is being removed in Vocal Mouse as it works for both speech and non-speech sound inputs. So, it can produce both discrete and continuous motion depending upon the user task requirement.

II. LITERATURE REVIEW

James R. Evans, Wayne A. Tjoland and Lloyd G. Allred [2] in 2000 developed a voice recognition and speech synthesis system. This system was reliable and speaker independent. Operator training was not required. Any 32-bit Windows software, which has window messaging capability, can access this voice control system. Standard programming languages such as Borland or Microsoft C/C+ + and Visual Basic (VB) and commercial packages such as Lotus Notes and Microsoft Word support this system. In this system, the operator was allowed to enter data and to control the software flow by voice command or from the keyboard or mouse. Grammar set, or legal set of commands could be specified dynamically. Recognition accuracy of the system got increased due to the use of reduced grammar set. The computer voice helps the operator to give his attention to other activities like probing a circuit card and taking readings. To insure reliable entry, system echoes the entered voice when operator is taking readings. Operator can hear the resultant reading using electronic tuning. This enables operator to focus on the circuit card instead of constantly turning his head to see the computer screen. They also described the capability and functionality of the Voice Control system.

Manuscript received on June 2012
Sandeep Kaur, Department of Computer Science, Lovely Professional University, Phagwara, India.

Susumu Harada, James A. Landay [3] in 2006 explained a system named Vocal Joystick for continuously controlling mouse pointer using vocal sounds. It was written in C++ and works on windows and Linux OSs. It needs no extra hardware other than sound card and microphone. In this paper they focused on the voice-based cursor control capabilities of the system. Vocal Joystick can act as a solution to the challenges faced by other mouse alternative devices. This system recognized sound based on its pitch and loudness. Its key benefit is that it processes the vocal input immediately and they are reflected at the interface without any kind of delay. The main aim of this project was to develop such a system which is easy to use, low-cost and takes input efficiently. They also presented a comparison between vocal joystick and two previous speech based mouse control systems to know novice user performance and preference. Other evaluations show that this system can be used by novice users without extensive training.

M. Abdeen, H. Mohammad, M. C.E. Yagoub [4] in 2008 presented a language-independent framework for a hands-free control of desktop computer. It works for PC windows and is based on DLL's (Dynamic Link Libraries).It is tested on both English and Arabic languages. The approach used for this framework is based on matching text produced by API functions and speech recognizer. After matching both the texts, the system runs the start menu. They also presented a comparison between their work and the XP and Vista. Some difficulties with speech recognition were encountered especially when users speak relatively above average. Reason for this is inability to recognize start and end of each word due to error in finding silences.

R. Maskeliunas, K. Ratkevicius, V. Rudzionis [5] in 2011 proposed a Voice-based human-machine interaction model for automated information services. This model allows recognize isolated commands together with some keywords. At the same model supports additional modalities such as OTA (Over the Air) menus of choices, controlled by touch and keyboard. Important characteristic of the model is the possibility to select a proposed choice (system of proposed selections). Different foreign language speech engines have different capabilities to recognize Lithuanian voice commands. Spanish speech engine enabled to achieve significantly higher recognition accuracy than English engine: overall recognition accuracy increased from 77% for the English engine to the 97% for the Spanish engine.

This research paper gives a new concept of Vocal Mouse as it works for both speech and non-speech sound inputs. So, it can produce both discrete and continuous motion depending upon the user task requirement. Details of this system are described in the remainder of this paper. The paper is organized as follows: Section III covers the objectives of vocal mouse and its significance, section IV covers the methodology of the system, section V covers implementation tool. Finally, conclusions and future scope are discussed in section VI.

III. THE VOCAL MOUSE

This research paper will describe the work that is done as a part of the Vocal Mouse project to use both words and vowel sounds as non-speech characteristics of human voice to enable fluid hands-free computing.

A. Objectives of Vocal Mouse

- Main objective of designing Vocal Mouse is to make interactions with existing computer applications possible or more accessible for users with motor impairments. This objective put more emphasis on the practical needs of physically disabled people [6] who can't use existing computer OSs and applications, which have been designed for keyboard and mouse input. Perhaps, Vocal Mouse may not provide the ideal solution from the interaction design perspective. Even then it will give benefit to people with motor impairments more than the alternative of not being able to access the functionality of computers at all.
- Another objective of the proposed system is to make interactions with existing computer applications more effective for general users. The main aim of this objective is to use voice input can as an additional input modality to augment the standard keyboard and mouse interaction for greater control i.e. giving voice input parallel to the keyboard and mouse. The integration of multiple input modalities permits greater expressiveness from complementary information sources and greater reliability due to redundancies across modalities [7].
- The other motivation behind this system is to design a new interfaces and application environments optimized for voice-based control. This objective seeks the ideal scenario in which voice input is given the highest priority from the input modality space, and focus is made to make the user interfaces more optimal.

B. System Overview

The Vocal Mouse will track both speech and non-speech vocal features including pitch, volume, and vowel quality [8] in real time using audio signal processing. This application will enable the user to control the mouse pointer smoothly and continuously by vocalizing various vowel sounds corresponding to the desired direction of movement. Under conventional speech-driven pointer control, only spoken words as commands are used. The rate at which such parameters can be changed is also limited by the speed at which each command phrase can be uttered and recognized by the speech recognizer. Movement direction is specified by vowel sound vocalizations, in which different vowel sounds are mapped to each of the four cardinal directions shown in figure 1. By varying the vowel sound and the volume continuously, the pointer's movement direction and speed can be smoothly controlled using the Vocal Mouse pointer control. This research paper will combine the functionality offered by the Vocal Mouse with traditional speech recognition engines to explore ways to extend the capability of voice-based interaction with new user interface technologies.

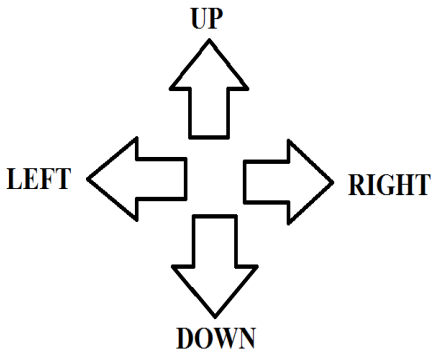

Fig 1: Four cardinal directions for mouse pointer.

C. Significance

The Vocal Mouse system has several key distinguishing features that provide benefits to its users:-

- VM (Vocal Mouse) system is based on the recognition of the words and vocal sounds which is a very robust and accurate method as compared recognition of words using conventional speech recognition systems. This concept is shown diagrammatically in figure 2.
- The instantaneous processing of every audio frame leads to much more immediate system response compared to systems that require a whole word or sequence of words to be recognized before an action is taken.
- Vocal characteristics such as vowel quality, volume and pitch can be changed by the user continuously, so the system allows such continuous changes to be transferred directly onto the control parameters, which results in smooth and responsive interaction.
- The only physical ability required of the user by the system is the ability to vocalize.
- It requires minimal equipment at very low cost. There is no need of expensive, bulky hardware for getting the inputs. The only thing that is needed is microphone.
- The proposed system can be used anywhere without need of any sort of supporting files and platform.
- It can be used anywhere because it is platform independent.
- This approach will be much helpful for the future purpose, in order to develop new applications with exciting and new features.

All these significant features of the proposed system make it suitable for the real time environment. The movement of the mouse cursor with the help of speech and non-speech vocal sounds makes it very much user friendly and flexible in nature.

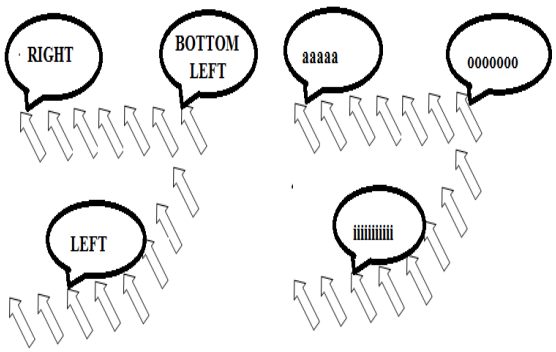

Fig 2: Vocal Mouse can use both speech (a) and non-speech vocal sounds (b) for pointer movement.

IV. METHODOLOGY OF CONSTRUCTING THE VOCAL MOUSE

Methodology of constructing the proposed system will consists of various modules. Each module uses different techniques and algorithms to perform its specific tasks. After a particular module completes its task, its output will become input for the next module. In the end the combined effort of each module will be displayed. Flowchart for various modules of the proposed system is shown in figure 3.

Module 1:- Acoustic signal processing
Module 2:- Pattern recognition
Module 3:- Motion control.

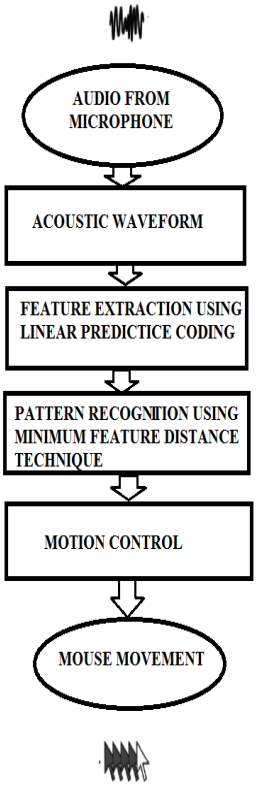

Fig 3: Flowchart for various phases of Vocal Mouse.

A. Signal Processing

The goal of the signal processing module is to extract low-level acoustic features using LPC technique that can be used in estimating the vocal characteristics. Process of feature extraction is shown in figure 4. The features to be extracted are:-

- Energy
- Normalized cross-correlation coefficients (NCCC)
- Gain

For feature extraction, the speech signal is PCM sampled at a rate of Fs=10,000Hz. Energy is measured on a frame-by-frame basis with a frame size of 25ms and a frame step of 10ms. Pitch is extracted with a frame size of 40ms and a frame step of 10ms. Multiple pattern recognition tasks may share the same acoustic features. Therefore, it is more efficient to separate feature extraction from pattern recognition. Working of LPC algorithm is shown in figure 5.

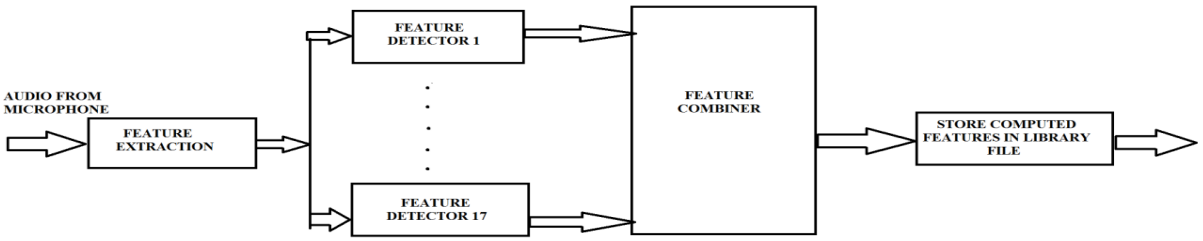

Fig 4: Process of Feature Extraction.

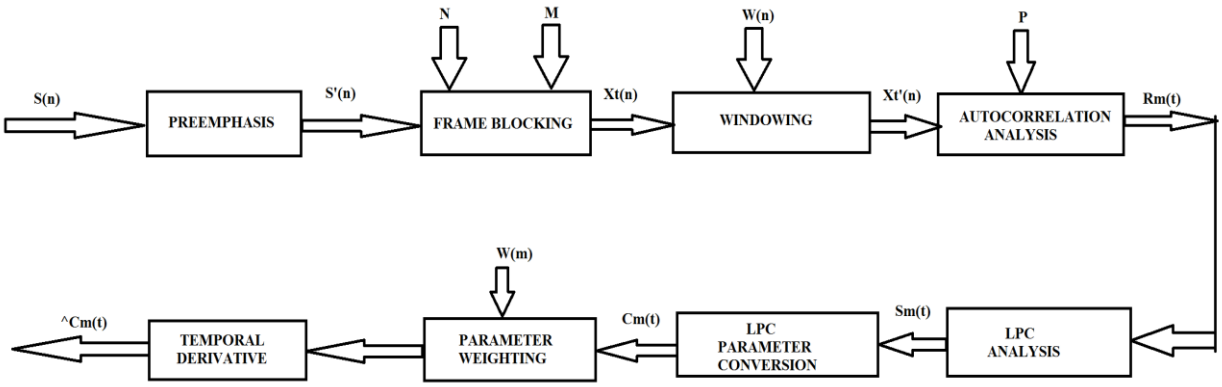

Fig 5: Block diagram of LPC (Linear Predictive Coding).

- The basic idea of LPC is that speech signal at time n, s(n), is the linear combination of past k signals .

  $S(n)=A_1.s(n-1) + A_2.s(n-2)+\ldots\ldots A_k.\ s(n\ldots\ldots\ldots(1)$

  Where A_1, A_2, $A_3$are constant coefficients

- $S(n)=\sum_{i=1}^{k} A_i.\ s(n-i) + G\ u(n)\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$

  where G is the gain of the excitation

  u(n) is the normalized excitation

- Expressing the above in z-domain, we get

  $S(z)= \sum_{i=1}^{k} A_i.\ z^{(-i)}.\ S(z) + G\ u(z)\ldots\ldots\ldots\ldots\ldots..(3)$

- Transfer function is

  $H(z) = S(z) / G\ u(z)$

  $= 1/(1-\sum A_i.\ z^{(-i)})$

  $= 1/A(z)\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(4)$

B. Pattern Recognition

This module uses the acoustic features to extract desired parameters. The estimation and classification system must simultaneously perform energy computation (available from the input), pitch tracking, vowel classification, and discrete sound recognition. Approach to be used is:-

Minimum Feature Distance Technique

This proposed technique is based on calculating distances between the spoken word and each word in the library shown in figure 6.

1) D=features of spoken word –features of word stored in library at training time.
2) Sum up all the corresponding differences.
3) Take the square root of the total calculated difference.
4) Perform the above step 1, 2 and 3 calculations for each command in library.
5) Above steps will result in four feature distance values (if number of commands stored in library are four)
6) Print the word with minimum feature distance.
7) The result will correspond to word spoken by user.

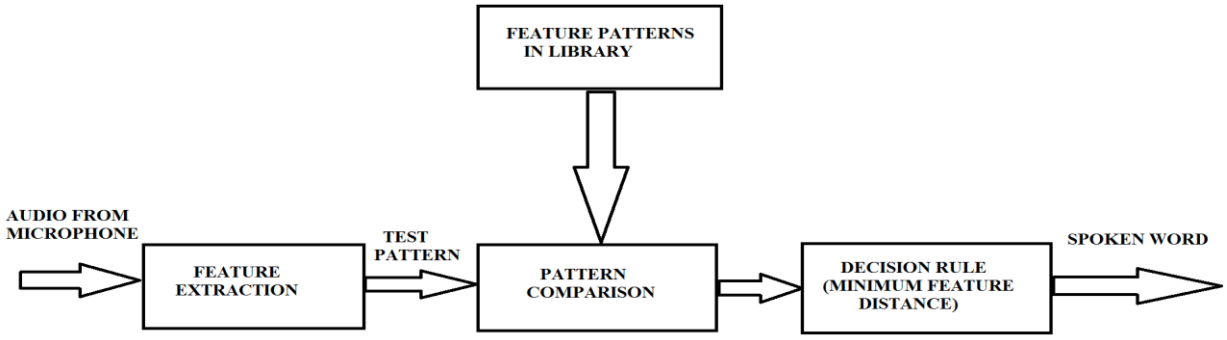

Fig 6: Process of Pattern Recognition.

Example showing the working of proposed algorithm

- Features of various commands stored in library in the form of a matrix:-
  Up→ $a_1, a_2, a_3, a_4 \ldots\ldots\ldots a_n$
  Down→ $b_1, b_2, b_3, b_4 \ldots\ldots\ldots b_n$
  Left→ $c_1, c_2, c_3, c_4 \ldots\ldots\ldots c_n$
  Right→ $d_1, d_2, d_3, d_4 \ldots\ldots\ldots d_n$
- Features of spoken word:-
  $s_1, s_2, s_3, s_4 \ldots\ldots\ldots s_n$
- Calculate corresponding differences
  $s_1-a_1, s_2-a_2, s_3-a_3, \ldots\ldots\ldots, s_n-a_n$
- Sum up all the corresponding differences.
  $\text{Sum}(s_1-a_1, s_2-a_2, s_3-a_3, \ldots\ldots\ldots, s_n-a_n)$
- Take the square root of the total calculated difference.
  $D1=\text{Sqrt}(\text{Sum}(s_1-a_1, s_2-a_2, s_3-a_3, \ldots\ldots\ldots, s_n-a_n))$
- Perform the above calculations for down, left and right also.
- Result will be a set of four feature distance values D_1, D_2, D_3, D_4 e g:- (17.6334,15.5128,7.5142,22.9444)
- Print the word with minimum feature distance. Word spoken by user is "left" because of its minimum feature distance as per taken sample values.

C. Motion Control: Direction and Velocity

This module receives several parameters from pattern recognition module and processes them to produce output more appropriate for determining cursor's 2-D movement. Relative energy instead of pitch will be used to control cursor velocity. It is observed that naturally users become quiet when they try to move cursor to small distance. Instead of comparing average energy level with global average energy, distribute the energy per frame among the different vowels according to the probabilities output given by the neural network and independently find out the average energy for each vocal sound.

D. Proposed Algorithm

After describing the full detail of working of VM the outlined algorithm for the system looks like:-

1) Call the microphone working verification module and repeat this step until user does not close the application
2) Call the noise reduction module [9]. Perform this step until the entire background noise is not getting removed.
3) Call the signal processing module and extract acoustic features like energy, MFCC etc. Perform this step only once.
4) Call pattern recognition module in which energy smoothing, pitch and format tracking and discrete sound recognition are performed. This module involves minimum feature distance technique.
5) Call motion control module to transform energy, pitch, vowel quality and discrete sound become acoustic parameters into direction, speed and other motion related parameters
6) Call application driver which will take motion control parameters to launch corresponding actions. End of algorithm.

V. IMPLEMENTATION TOOL ANS RESULTS

The proposed system can be implemented using MATLAB. MATLAB is a programming environment for algorithm development, data analysis, visualization, and numerical computation. To do work on speech recognition, users can take the help of VOICEBOX. VOICEBOX is a speech processing toolbox consists of MATLAB routines. There is no need of any other expensive, bulky hardware. The only thing that is needed is good quality microphone.

A. Training Phase

Vocal Mouse project starts with training phase. Get four words (up, down, left, right) or four sounds (aaaaa, iiiiii, eeeeee, oooooo and uuuuuuu) from microphone and compute their features. Firstly, save the calculated features in a feature matrix and then store them in some other file. User is given 1 second to say each word. User will press enter and say the specified word in 1 second. Features of all the spoken words are stored in a feature matrix fw.

Matrix fw is a 2-D matrix with 4 rows (one for each word) and 17 columns (17 features are extracted). Contents of this matrix can be loaded to library data file for further reference. Here is the completion of training phase.

B. Testing Phase

During the testing phase, User is given 1 second to say any word. User will press enter and say the needed command word in 1 second. Features of the spoken words are computed and these features are compared with feature patterns already stored in library during training time. For this comparison, minimum feature distance approach is used. Vocal Mouse system will compute resulting spoken word on the basis of minimum distance. Output will be displayed on the console window of MATLAB giving "spoken word is:…" along with FFT (Fast Fourier Transform) of spoken command as shown in figure 7. After the command recognition, system will perform task as per the requirement of the user. When spoken word is "left", mouse will start moving in the left direction as shown in figure 8.

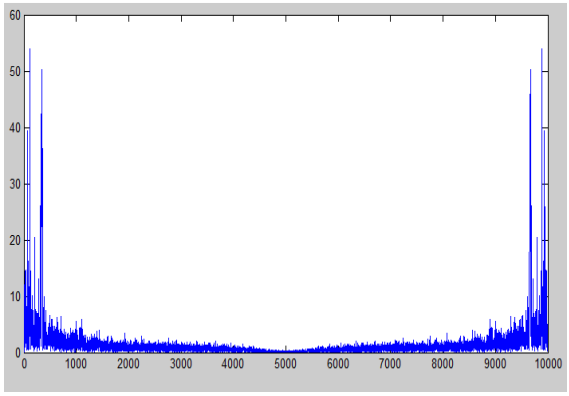

Fig 7: Fast Fourier Transform of spoken word.

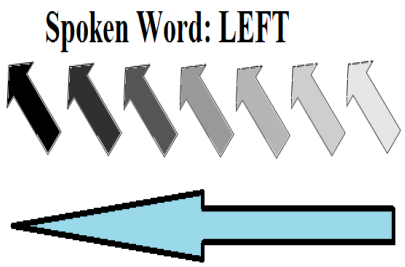

Fig 8: Resultant mouse movement with command "left".

VI. CONCLUSION AND FUTURE WORK

A. Points Concluded

More interactive and easy to use:- Vocal Mouse is a system that enables a user to continuously control the mouse cursor using their voice. Vocal Mouse used a new technique "minimum feature distance algorithm" for pattern recognition. It has two primary operating modes: the speech mode and the vocal mode. Various evaluations indicate that Vocal Mouse provides more interactive and easy to use interface as compared to previous cursor control techniques. Users feel more comfortable to work in two optional modes. During the speech input mode, the user can issue any of the standard command to control the computer. When the user wishes to start controlling the mouse pointer, the command "switch" switches the user to the vocal mode. During the vocal mode, the user can use the vocal sounds to fluidly manipulate the mouse pointer. Evaluations with novice users revealed that the Vocal Mouse can be effectively operated and that its performance beats the existing speech-based cursor control methods.

Minimum training required:- Novice users of the Vocal Mouse pointer control application need minimal training of five minutes to perform basic tasks. An evaluation is conducted which involves ten participants. Each participant is given five minutes of Vocal Mouse training. Results show that the average target acquisition times for the Vocal Mouse pointer control method is roughly three times faster than the Mouse Grid technique.

B.Future Work

Implementation of the Vocal Mouse can be continued further and various kinds of analysis can be performed by recruiting participants, in particular those with various motor impairments, to try out our system. In future, vocal mouse can be used to develop applications such as drawing and games. These applications play an important role in enriching the lives of people especially those whose range of activities may be limited due to some disability. Vocal Mouse is well suited for such applications, and will be exploring ways in which the system can be best used to support them. Future planned improvements in the algorithms underlying the Vocal Mouse (to improve accuracy, user-independence, adaptation, and speed) will further increase the VM system's viability, and combined with practice could improve VM enough so that it becomes a reasonable alternative compared to a standard mouse's performance.

C. Enhancing Vocal Mouse Application

As more users express interest in the Voice Mouse, a number of enhancements need to be made to ensure that it offers as intuitive and effective solution as possible.

- One area of enhancement is in the visual feedback and the user interface for supporting self-diagnosis of recognition issues. A testing mode can be provided for the user to test whether or not the system is responding properly. More information should be conveyed to the user that would allow them to troubleshoot instances when the system fails to recognize certain inputs.
- A method for automatically coaching the user to improve their vocalization, especially of the vowel sounds, will be beneficial. New module can be added to Voice Mouse which will provide video samples of each sound, real-time feedback during the testing mode as well as explicit suggestions to the user regarding how to change their mouth shape or other features to approach the desired sound.

• The main obstacle and sources of frustration that a user faces is "false positives". False positive means system generates some recognition event when the user did not intend to vocalize. This could happen either when the user forgets that the system is processing vocal input and begins speaking or making some sounds, typically when the user's attention is away from the user interface, or when the system picks up some background noise and incorrectly recognizes it as some valid vocalization. In both cases, the user may not realize that the system has processed the false positive events until sometime later, e.g., when the user turns his attention back to the interface, at which point the user may become confused about what had happened, and possibly quite frustrated about not knowing what to undo if the exact series of actions that were inadvertently executed is not immediately apparent. In such a situation, quick method is required to disable current processing. A possibility for future work for addressing this issue is the use of various external contexts such as the user's "head posture" and "gaze" to disable voice input when the user is likely disengaged from the interface.

REFERENCES

1. Susumu Harada, "Harnessing the Capacity of the Human Voice for Fluidly Controlling Computer Interface", University of Washington, 2010
2. James R Evans, Wayne A Tjoland and Lloyd G Allred, "Achieving a Hands-Free Computer Interface using Voice Recognition and Speech Synthesis ",IEEE AES Systems Magazine, 2000
3. Susumu Harada and James A Landay, "The Vocal Joystick: Evaluation of Voice-based Cursor Control Techniques", Portland, Oregon, USA: ASSETS, 2006
4. M Abdeen, H Moshammad and M C E Yagoub, "An Architecture for Multi-Lingual Hands Free Desktop Control System for PC Windows", Niagara Falls, Canada : IEEE , 2008
5. R Maskeliunas, K Ratkevicius and V Rudzionis, "Voice-based Human-Machine Interaction Modeling for Automated Information Services", ISSN 1392-1215 Electronics and Electrical Engineering, 2011
6. R Norma Conn and Michael McTear, "Speech Technology: A Solution for People with Disabilities", Savoy Place, London WCPR OBL, UK: IEE, 2000
7. Minh TU Vo and Alex Waibel "A Multi-Lingual Human-Computer Interface: Combination of Gesture and Speech Recognition", Carnegie Mellon University Pittsburgh, U.S.A, 2009
8. Susumu Harada, Jacob O Wobbrock, Jonathan Malkin, Jeff A Bilmes and James A Landay ,"Longitudinal Study of People Learning to Use Continuous Voice-Based Cursor Control", Boston, MA: Conf. on Human Factors in Computing Systems
9. M Rahmani, N Yousefian and A Akbari, "Energy-based speech enhancement technique for hands-free communication", ELECTRONICS LETTERS Vol. 45 No. 1, 2009

AUTHORS PROFILE

Sandeep Kaur She is an Assistant Professor at Lovely Professional University, Phagwara, India. She has completed her B-Tech (H)-M-Tech (CSE) from Lovely Professional University. Her M-Tech specialization is Internet Technologies. Her areas of interest are: Artificial Intelligence, Speech Recognition and Intelligent Systems.