Project Report
ON
"SPOCK: AN ADVANCED HUMAN COMPUTER INTERACTION SYSTEM"

(A Project Report submitted in partial fulfillment of the requirements of Bachelor of Technology (Hons.) in Department of Computer Science and Engineering at National Institute of Technology, Jamshedpur)

Submitted by:

Sonal Raj (487/10)
Shashi Kant (462/10)

Under the guidance of

Mr. Sanjay Kumar
Associate Professor
Dept. of Computer Science and Engineering

NATIONAL INSTITUTE OF TECHNOLOGY

(Formerly Regional Institute of Technology, Jamshedpur)
JAMSHEDPUR – 831014
(Deemed University)

Director: 0657-2407614 (O)
☎ : 2407598, 2407642 (O)

Email: director@nitjsr.ac.in
FAX: 0657-2382246

NATIONAL INSTITUTE OF TECHNOLOGY

JAMSHEDPUR – 831014

*Ref. No. **NIT/***

*Date- **07/05/201**4*

TO WHOM IT MAY CONCERN

This is to certify that the Project entitled **"An advanced Human Computer Interaction System"** carried out by **Mr. Sonal Raj, Registration No-CS110487, Mr. Shashikant Kumar, Registration No-CS110462,** bona fide students of **National Institute of Technology, Jamshedpur** in partial fulfillment for the award of **Bachelor of Technology (Hons.) in Computer Science & Engineering,** during academic year 2013-2014 has been successfully completed under my supervision and guidance at National Institute of Technology, Jamshedpur. They have fulfilled the requirements laid down for the Project Work.

(Mr. Sanjay Kumar)
Project Guide
Associate Professor
Dept. of Computer Sc. & Engg.

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose constant encouragement has been a source of inspiration throughout the course of the project.

We express our sincere gratitude to **Prof. RamBabu Kodali**, Director, National Institute of Technology, Jamshedpur for providing us an opportunity to undertake a project as a part of our curriculum.

We express our gratitude to **Mr. R. R. Suman,** Head of the Department, Computer Science and Engineering, NIT Jamshedpur for his valuable support.

We gratefully acknowledge the guidance provided by our project guide **Mr. Sanjay Kumar**, Associate Professor, Department of Computer Science & Engineering, NIT Jamshedpur. He had the kindness to accept being our Project Guide and let us work with autonomy. Discussing with him the technical knowhow of our project was really helpful.

We are highly obliged to all other faculty members and staffs of department for the supportive and cordial demeanor exhibited towards us. We thank them for their technical assistance and concerned documents provided by them.

Sonal Raj (487/10) **Shashi Kant (462/10)**

ABSTRACT

Having been avid users of popular general purpose Linux distros for the past few years we noticed that some major forms of human – computer interactions was negligibly developed for such type of systems. Voice activation, command control and Text-to-Speech were major features either lacking completely or rather under-developed which in a way produced limited operability for common end-users or physically-challenged users.

In this project, we work on a speech recognition and voice – control engine initially inherent to the UNIX environment. This proposed engine would serve as a framework for several command tools for purposes including File Management, User operations of the OS, Basic Dictation for Word processors or similar software, Web Browsing, etc. The above functionalities are introductory uses for the proposed engine.

This would provide a more elegant way of Human-Computer Interactions on a Unix Platform thereby increasing the User Experience. Moreover, the tools developed with this could enable physically challenged users to manage their operations more easily.

We have based our work mostly in C/C++, Python, JavaScript and Java. The platform used is UNIX. We use OpenCV with encoded custom filters, and PocketSphinx improving the accuracy compared to existing alternatives. The Speech Recognition algorithms and bindings will be based on the work done under the Sphinx project at Carnegie Mellon University. Other tools will be updated as and when required in the future builds.

The project holds great prospects and is definitely a large one, but, we present a working minimal implementation prototype as our BTP work. The code is Open Source and is released under the MIT License.

CONTENTS

INTRODUCTION

Contrast enhancement techniques are widely used for image/video processing to achieve wider dynamic range. Histogram modification based algorithm is the most popular approaches to achieve widely dynamic range with more clear density value.

1.1 Objective

This project is basically constituted of two distinct, but interdependent zones of features. The first zone includes the development of the Human Computer Interaction Engine which would be an integrable extension to most computing systems providing an advanced alterative to the present way of communicating. Secondly, the project also aims to present an Application Programming Interface to developers of third party applications resident on those systems, either through language specific methods or through RESTful APIs, in order to make use of this engine for providing advanced interaction techniques to their applications.
Thus the Prime Objectives of the Project can be highlighted as follows:

1. Obtain a viable knowledge of the working of existing algorithms and methods of interaction, and build up a concept system to improve these techniques.

2. Code a Human Computer Interaction Engine capable to techniques like motion and gesture recognition, speech synthesis (biometric authentications, iris control and many other features could be included in future versions and are beyond the scope of this project.).

Implement an Application Programming Interface (API) to extend the functionality of this engine to multitude of applications running on the system to implement it in their own way.

1.2 Motivation

The motivation for taking up this project is definitely not compulsion or marks! It the underlying enthusiasm in the minds of the team members to create something useful with the knowledge gained in the years of engineering, to make a contribution by being the flag-bearers of technology and a dream to make a dent in the universe.

Inspired by brilliant pieces of work like "The Sixth Sense" by Pranav Mistry (Massachusetts Institute of Technology) which revolutionized the era of touch free interaction with the machine world, or the epics of modern sci-fi like Star Trek or the Jarvis interactive computer in the Iron Man movie series, or Microsoft Kinect gaming system, we decided to take a step towards building this concept. After all, the fiction of today is the reality of tomorrow! There is a huge scope of evolution and improvement in the field of interactive computing and we are planning to tap into some of its areas through this project.

Pranav Mistry's work at MIT on Sixth Sense can be checked out at: http://www.pranavmistry.com/projects/sixthsense/

Iron Man's J.A.R.V.I.S computer system can be referenced at : http://marvel-movies.wikia.com/wiki/J.A.R.V.I.S.

1.3 Mission statement

This Project's mission is to develop a technological product which will change the way you interacted with your computers using life-like sensing methods.

SURVEY OF EXISTING WORK

Since this project will currently take into consideration the gesture and speech synthesis parts of an advanced interaction system, the survey areas have been more emphasized on these aspects.

Projects

1. **Sixth Sense**: 'SixthSense' is a wearable gestural interface that augments the physical world around us with digital information and lets us use natural hand gestures to interact with that information. It was built at MIT Media Lab in 1997 that combined cameras and illumination systems for interactive photographic art, and also included gesture recognition (e.g. finger-tracking using colored tape on the fingers)
2. **Microsoft Kinect:** Kinect (codenamed in development as Project Natal) is a line of motion sensing input devices by Microsoft for Xbox 360 and Xbox One video game consoles and Windows PCs. Based around a webcam-style add-on peripheral, it enables users to control and interact with their console/computer without the need for a game controller, through a natural user interface using gestures and spoken commands.
3. **Sphinx:** CMUSphinx toolkit is a leading speech recognition toolkit with various tools used to build speech applications. CMU Sphinx toolkit has a number of packages for different tasks and applications. Things like building a phonetic model capable of handling an infinite vocabulary, post processing of the decoding result, sense extraction and other semantic tools are missing and need to be worked upon.
4. **SIRI and Google Voice Search:** SIRI is an intelligent personal assistant and knowledge navigator which works as an application for Apple Inc.'s iOS. The application uses a natural language user interface to answer questions, make recommendations, and perform actions by delegating requests to a set of Web services. Apple claims that the software adapts to the user's individual preferences over time and personalizes results. Google's Voice Search is a similar adaptation.

Papers and Publications

1. **Real-time hand gesture recognition using range cameras,** *Hervé Lahamy and Derek Litchi,* Department of Geomatics Engineering, University of Calgary, NW, Calgary, Alberta

[The proposed methodology involves 3D image acquisition, segmentation of the hand information, tracking of the hand blob, recognition of the hand gestures and determination of the position and orientation of the user's hand. This research demonstrates the capability of a range camera for real-time gesture recognition applications**.]**

2. **Real-Time Human Pose Recognition in Parts from Single Depth Images,** *Jamie Shotton Andrew Fitzgibbon,* Microsoft Research Cambridge & Xbox Incubation.

   [This paper proposes a new method to quickly and accurately predict 3D positions of body joints from a single depth image, using no temporal information. It takes an object recognition approach, designing an intermediate body parts representation that maps the difficult pose estimation problem into a simpler per-pixel classification problem. **]**

3. **Minimum variance modulation filter for robust speech recognition**. *Yu-Hsiang Bosco Chiu and Richard M Stern,* Carnegie Mellon University, Pittsburgh, USA.

   [This paper describes a way of designing modulation filter by data driven analysis which improves the performance of automatic speech recognition systems that operate in real environments. The filter for each nonlinear channel output is obtained by a constrained optimization process which jointly minimizes the environmental distortion as well as the distortion caused by the filter itself. Recognition accuracy is measured using the CMU SPHINX-III speech recognition system, and the DARPA Resource Management and Wall Street Journal speech corpus for training and testing. **]**

4. **Some recent research work at LIUM based on the use of CMU Sphinx***, Yannick Estève ET. Al.* LIUM*,* University of Le Mans, France

   [This paper presents an overview of the recent research work developed at LIUM using the CMU Sphinx tools. First, it describes the LIUM ASR system which reached very competitive results on French evaluation campaigns. **]**

Several other online resources including Software and Project Documentations of the above mentioned Projects as well as blogs and wikis have also been referenced to for more insights into the working principles and module integration ideas.

SOFTWARE SPECIFICATIONS

Proposed Work

With information gained from the above surveys and studies, we propose a multi-modal Natural User Interface (NUI) for personal computing devices to work as a human-computer interaction engine. The framework would initially integrate gesture and speech synthesis for common applied tasks and would be extended to biometric, motion and other standards in future aspects. The proposed work would not only be useful to an everyday computer user but also act as a radical instrument of change for the disabled and physically challenged. Moreover, we also propose to make the work of this project open-source and provide APIs to extend its functionality to several artificial intelligence and search operations. The project will make use of several algorithms optimized in performance to create a near-real time experience of the synthesis process, thereby being used as a viable resource for time-critical scenarios like threat analysis in the national security requirements.

Technologies used

The list of technologies to use is tentative, and can be varied based on the enhanced needs or failure alternative use.

1. **Languages**
    a. **C++**: We have written most of the modules for both speech and gesture synthesis in C/C++ since these operations make use of the system/kernel level tools, which are quite effective and fast in these languages.
    b. **Python**: This language has been used mostly for scripting purposes in order to perform complex stream processing or handling large and complicated data structures.
    c. **Bash**: This language will be used for shell scripts on a typical UNIX system running the bash shell so that the engine and its APIs are available to the end user easily.
2. **Frameworks**
    a. **OpenCV**: OpenCV (Open Source Computer Vision Library) is a library of programming functions mainly aimed at real-time computer vision, developed by

Intel, and now supported by Willow Garage and Itseez. It is free for use under the open source BSD license. The library is cross-platform. It focuses mainly on real-time image processing. If the library finds Intel's Integrated Performance Primitives on the system, it will use these proprietary optimized routines to accelerate itself. The image processing used for the gesture recognition system will be formed with the real-time optimized libraries of OpenCV.

b. **CMUSphinx**: CMUSphinx toolkit is a leading speech recognition toolkit with various tools used to build speech applications. CMU Sphinx toolkit has a number of packages for different tasks and applications. Things like building a phonetic model capable of handling an infinite vocabulary, post processing of the decoding result, sense extraction and other semantic tools are missing and need to be worked upon. We will be using the Sphinx additives with a custom made dictionary for the swift recognition of the input voice stream.

MAJOR AREAS OF IMPACT

The work of these project has profound influences in the tech community with its impacts ranging from personal computing experience to national security. A NUI is implementable in every area. The major areas and impact of this project on these areas can be outlined as follows:

1. The project will bring a more lively experience to the “muggles” of the computing world in interacting with the system. It is more natural for people to learn the use speech and gestures rather than learning to type!
2. Spock will lead to more robust devices which are equipped with few sensors and capable of processing crucial data.
3. An application of the Spock API could be used for “Video Search” Techniques. In general keyword or term based searches are prevalent. But what if the search query is available in the middle of a playing video. With right kind of systems and use of the Spock API, we could extract intra video information to make the search more precise and elaborate.
4. The Spock API also has profound impact (of course, with future modifications) to the fields of National Security, where monitoring data could be analyzed for sequences of threat related information.

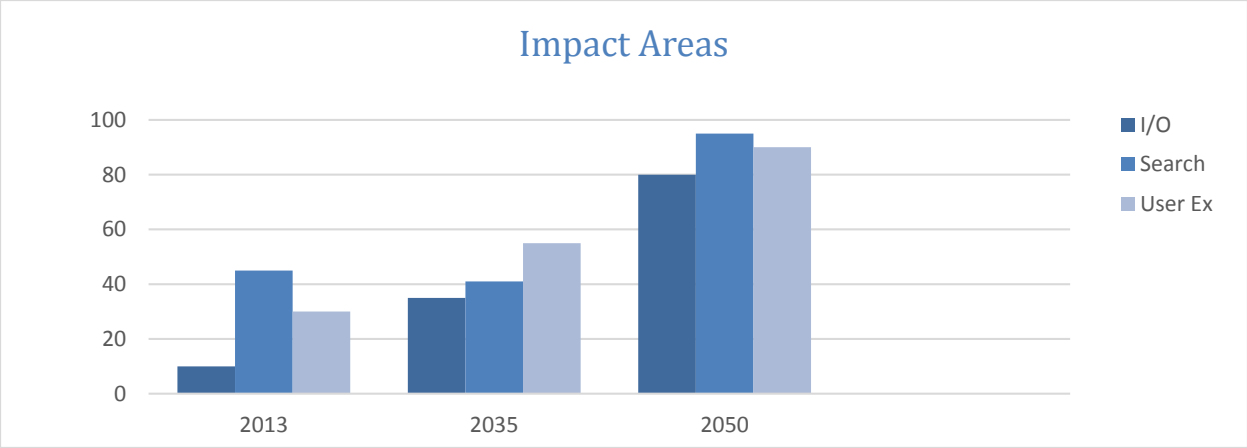

Fig: Impact areas and their proposed impact percentages

Why are companies like Microsoft and Intel working hard to lay their hands on an effective NUI for the common man?

For an adult with no prior exposure or experience, learning how to use a desktop computer can be a confusing challenge.

The desktop computing experience is neither intuitive nor innate to human beings – it requires significant training, time, and ideally early-age immersion in order to understand the paradigms of computing (both understanding how to physically interact, as well as conceptually understand virtual computing environments).

Since the adoption of the personal desktop computer in the 1980s, our efforts to naturalize the personal computing experience has been limited to humans as the variable factor when adapting to computing environments. It's amazing that we have still managed to retain the paradigms of how to sit at a computer since the early 80s – with desktop computing today resembling the exact same monitor/keyboard/mouse setup with very little physical interactive variation. To illustrate how unnatural and unintuitive this archaic experience is, imagine the learning curve that a first-time user in their 50s experiences in order to understand this interface paradigm. There really is nothing fundamentally "natural" about the desktop computing experience – if anything, it is the furthest thing away from being a natural human function.

We are living in a fascinating time – it is only recently that we are finally breaking the paradigms of our traditional interface constraints set by the 80s. With the introduction of the biggest paradigm shift in human-computer interfaces, we are finally seeing new forms of portable computing devices – multitouch surfaces, powerful and lightweight mobile devices, and now the emerging market of wearable technology. We are entering an exciting world outside the constraints of physical and virtual environments. It's about time we return to our natural world.

We are now moving from touch to no-touch. It is easier (and hygienic too!).As Samsung would have quoted "designed for Humans".

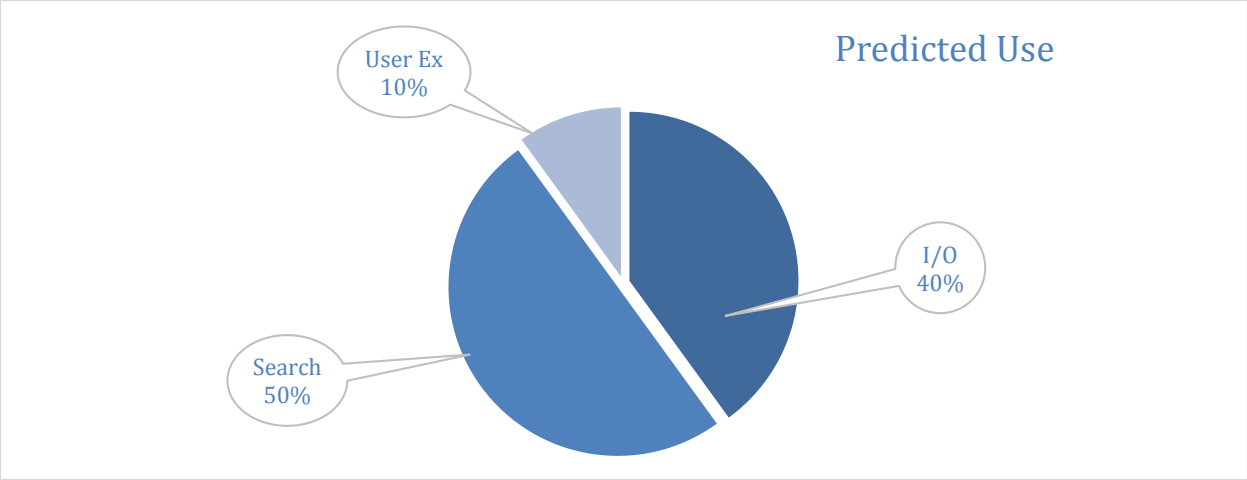

Fig: Predicted Use of a Spock Engine and API in a decade after initiation.

The Spock Project is targeted to affect the following areas in a major may, as is evident from the statistics presented above:

1. **Search**: Looking up information is the highest rated activated among users today, and it boils down to an efficient search system. Using of Spock can help to implement better semantics for search queries. This technique also makes it faster to extract the query since most people talk faster than they can type. Huh, humans!
2. **Input/Output**: This is a more in demand area that hold the future of technological gadgets and communication devices. In developing countries, where a majority of the population cannot directly interact with computes can now utilize a more efficient way to communicate which does not require much technical skills.
3. **User Ex**.: This paves the way for next generation computing devices which will get rid of the unnecessary peripherals which contribute to the bulkiness of the device. Thus future of devices will be more sleek, and user friendly options.

PROGRESS REPORT

Technologies learnt and configured

So far we have studied documentations of the OpenCV Framework and the Sphinx Speech toolkit from CMU and fiddled around with them quite a bit. Thereby, their configurations in Linux as well as Windows was carried out and learnt successfully.

It was an exciting experience to see the our computing devices respond naturally to human gestures and speech and it is intriguing to learn that self-learning and training module could be developed for our systems as well. We also had quite a hands on experience with Linux Kernel Integration methods as well as learnt a lot about speech synthesis with Python and CMU Sphinx.

We also dealt with JavaScript for our API creation and its use with the Web Framework demos and even tweaked them for higher efficiency. The JavaScript API implementations of OpenCV and Sphinx can now be used to develop expert interactive web applications in collaboration with the browser features like Google Chrome and Firefox extensions.

Coding Work

We began working on the codes of several modules based on the respective frameworks. The code created till date are related to the following functionality.

a. Gesture Registration framework and Database Creation, in which, the Spock application can now register human gestures as an input and store them in an indexed format for the mapping with specific keys and combinations.
b. Noise Reduction and image synthesis: This is done by subtracting the RGB value of the pixels of the previous frame from the RGB values of the pixels of the current frame. Then this image is converted to octachrome (8 colors only - red, blue, green, cyan, magenta, yellow, white, black). This makes most of the pixels neutral or grey. This is followed by the greying of those pixels not surrounded by 20 non-grey pixels, in the function crosshair (IplImage*img1, IplImage* img2). The non-grey pixels that remain represent proper motion and noise is eliminated. A database is provided with the code which contains a set of points for each gesture.

c. Active Gesture Adaptation: This has been implemented in its most primitive form and all that it does it to record the smallest deviations from a standard stored gesture and match it with closest available gesture in the database.
d. Speech synthesis: Split the waveform by on utterances by silences match all possible combination of words with the audio models used in speech recognition acoustic model phonetic dictionary Language Model.
e. Web Integration: The same modules have been translated into the form of a JavaScript Library which can now be used in web applications to use the client's hardware like cameras and microphones and process the input in the client side itself to produce web gestures to interact with the websites without any physical communication.

The Progress of the coding process can be followed at the Github repository: https://github.com/sonal-raj/Spock/

In the Hindsight

Keeping in mind the potential of the Spock Project as a whole, it is necessary to take a look at what prospects holds for the project in the near future and all that it can achieve in the future rounds of coding works from the community, being an open source project.

a. **Activation and Pacification:** This feature is one of the most important to be implemented aspects of the Spock Project. This aspect deals with the system being able to anticipate when to listen and see the relevant gestures in spite of being active all the time and when to ignore casual conversations or gestures in the domain of definition.
b. **Recognize Patterns:** The Spock Project can be harnessed to detect activities based on certain user specified patterns and hence it can include modules on machine learning and streamed image processing, which will enable it to be used in defense systems and other security aspects at both small and large scales.
c. **Application Integration:** When the system is fully developed, being open source, Application developers for all platforms can now use this in their applications and ship a bundled copy which being local in nature can help to integrate better in an application specific manner.
d. **Streaming API:** This is the cloud based version of the API which will be serviced through a RESTFUL Interface and it can be directly used by applications on mobile devices and other network connected interfaces which run on low resources and can now use the functionality through a simple request to the cloud hosted application.

CONCLUSION

This project presents a minimalistic framework for an advanced human computer interaction system which can now be used interact any version of the computing devices with basic hardware installed to provide an amazing experience of natural interaction with the system as a whole.

This system rings together several small open source projects along with custom developed bindings and adapters in order to achieve a unified and more responsive system to be used by common people as end-users. Hence, it implements an abstraction to non-developers, while providing an open source alternative to the current systems of interacting with our personal devices.

As already stated, this Spock Project holds immense future possibilities, even the scope of becoming an industry milestone. And our work for this BTP project lays the foundation of this project, beginning the series of impacts to the future that are yet to come. We have released our work on an open source license, so that other developers can now work upon it as well and coming together as a developer community to make this product viable and operable to the end user as much as possible.

REFERENCES

Projects

1. Sixth Sense, Pranav Mistry, MIT Labs
2. Microsoft Kinect, Microsoft Research and XBox
3. CMUSphinx, Carnegie Mellon university
4. SIRI and Google Voice Search, Apple and Google, Inc.

Papers and Publications.

1. **Real-time hand gesture recognition using range cameras,** *Hervé Lahamy and Derek Litchi,* Department of Geomatics Engineering, University of Calgary, NW, Calgary, Alberta
2. **Real-Time Human Pose Recognition in Parts from Single Depth Images,** *Jamie Shotton Andrew Fitzgibbon,* Microsoft Research Cambridge & Xbox Incubation.
3. **Minimum variance modulation filter for robust speech recognition**. *Yu-Hsiang Bosco Chiu and Richard M Stern,* Carnegie Mellon University, Pittsburgh, USA.
4. **Some recent research work at LIUM based on the use of CMU Sphinx***, Yannick Estève ET. Al.* LIUM*,* University of Le Mans, France