A Thesis

entitled

Multimodal Data Fusion Using Voice and Electromyography Data for Robotic Control

by

Tauheed Khan Mohd

Submitted to the Graduate Faculty as partial fulfillment of the requirements for the Doctor of Philosophy Degree in Engineering

Dr. Ahmad Y. Javaid, Committee Chair

Dr. Mansoor Alam, Committee Co-Chair

Dr. Devinder Kaur, Committee Member

Dr. Weiqing Sun, Committee Member

Dr. Jared Oluoch, Committee Member

Dr. Quamar Niyaz, Committee Member

Dr. Cyndee Gruden, Dean
College of Graduate Studies

The University of Toledo
August 2019

An Abstract of

Multimodal Data Fusion Using Voice and Electromyography Data for Robotic Control

by

Tauheed Khan Mohd

Submitted to the Graduate Faculty as partial fulfillment of the requirements for the Doctor of Philosophy Degree in Engineering

The University of Toledo
August 2019

Wearable electronic equipment is continuously evolving and is increasing the human-machine integration. While industrialists (read Elon Musk) want to integrate a microchip in the human brain to leverage the faster processing capabilities of machines, others have been trying to build human-like machines. Available in various forms, these sensors can detect and measure the physiological changes in the human body; and may use those signals to control other devices. One such sensor, an electromyographic sensor (EMG), captures electromyographic data using myoelectric (electric signals in muscles) signals and translates them to be used as input signals through pre-defined gestures. Use of such a sensor in a multimodal environment will not only increase the possible types of work that can be accomplished with the help of such a device, but it will also help in improving the accuracy of the tasks performed. This research addresses the fusion of input modalities such as speech and myoelectric signals captured through a microphone and EMG sensor, respectively, to accurately control a robotic arm. The research was completed in three phases.

During phase 1, an extensive survey on technologies based on the multimodal environment was conducted. The goal was to find the pros and cons of each application and its utility. The classification was broadly divided into unimodal and multimodal systems. The multimodal system was further classified based on the fusion of input

modalities. Phase 1 results reaffirmed our expectation that the EMG data along with speech has not been used in many multimodal systems and if used, hasn't resulted in a high accuracy fusion that is useful for real-world application.

Phase 2 involved performing the experimental research using the EMG data (collected using the EMG sensor) with speech (collected using a speech recognition API). The findings show that there is a scope of improvement in accuracy for both the modalities when the EMG and Speech data was collected in laboratory conditions. The error percentage for the modalities varies from 8.9-34.1%. A decision-based fusion was performed which lead to a conclusion that multimodality improves the accuracy of operating the robotic arm, and the error rate reduced to 3.5-7.5%.

The last phase dealt with improving the results achieved during phase 2 using machine learning techniques and analyzes the most suitable strategy for controlling a robotic arm. Six machine learning algorithms were tested. After the training data was provided with sixty error conditions and tested again on the newly developed cases, the highest accuracy achieved through the K-nearest neighbor (KNN) algorithm was approximately 92%. Building upon phase 2, phase 3 concluded that use of machine learning algorithm helps in cases where input is misinterpreted and the error reduces drastically.

Dedicated to my Parents

Mr. Mohd Tahir Khan and *Mrs. Hamida Begum*,

my elder brother, *Mr. Mohd Tauqeer Khan*,

my wife, *Mrs. Sadiqa Sumbul*,

and my kids *Hamad* and *Ibaad*

for their never-ending love, support, trust, and patience

Acknowledgments

I would like to express my deepest gratitude and appreciation for my adviser Dr. Ahmad Y. Javaid, for all his guidance, leadership, patience, and support during the course of my graduate studies. He not only provided me with excellent environment and facilities for my research but also helped me become a better person through his constant support at times when I was struggling to stay focused and productive.

A special thanks to my committee member, Dr. Quamar Niyaz, whose contribution in designing the algorithms helped me carry out my research in a smooth manner.

I would also like to take this opportunity to specially thank my committee member, the EECS Chairman, Dr. Mansoor Alam, for overall support, guidance, trust and granting me with the opportunity to pursue my master's degree & PhD at The University of Toledo.

I would like to acknowledge the EECS department at the University of Toledo for giving me a tuition scholarship that helped me support myself financially. I would also like to extend my appreciation to the University of Toledo for providing me a platform to chase my dream and passion.

I would also like to express my appreciation to the Center for Creative Instruction for providing me with the funding for the Graduate Assistant-ship. A special thanks to William McCreary, vice president, CIO, and CTO for providing me an opportunity to work as a Graduate Assistant. I cannot thank my parents enough for everything that they have done for me. I would also like to thank my wife, Sadiqa, for being a source of motivation for me. Finally, I would also like to acknowledge the

support of my older brother and my friends.

Contents

List of Tables

List of Figures

List of Abbreviations

AI	Artificial Intelligence
API	Application Programming Interface
ATM	Automated Teller Machine
AVS	Amazon Voice Service
CRT	Cathode Ray Tube
CSV	Comma Separated Values
EMG	Electromyography
GPS	Global Positioning System
GUI	Graphical User Interface
HCI	Human Computer Interaction
HMI	Human Machine Interaction
LCD	Liquid Crystal Display
LED	Light Emitting Diode
ML	Machine Learning
MMDF	Multi Modal Data Fusion
PDA	Personal Digital Assistant
PUI	Perceptual User Interface
SSI	Social Signal Interpretation
SVM	Support Vector Machine
UI	User Interface
USB	Universal Serial Bus
VR	Virtual Reality

List of Symbols

σ .... Standard deviation in the error percentage after fusion of modalities

μ .... Mean of the error percentage after fusion of modalities

$f(x)$ .... Function of x

$\sum$ .... Summation

Chapter 1

Introduction

Human-machine interaction (HMI) seeks to enable machine adaptation to the application needs and involves the study of interfaces that facilitate the synergy between people and a machine. Also known as human-computer interaction (HCI), HMI primarily classifies interactions as unimodal and multimodal based on the number of modalities used in communication for interaction. Unimodal interface examples include a touch, speech, or gesture-based user interface [1]. Multimodal HMI combines natural input methods such as speech, touch, gestures, pen, or body movements in a synchronized manner and then transforming it to a multimedia system output. These systems make use of novel technologies for drawing I/O and allow the users to perform tasks with higher accuracy and precision despite any physical constraints or limitations [2]. The application of the multimodal systems is widely popular in robotics, fixed or mobile, and serves as one of the primary motivations for research in this field with the goal of developing a user-friendly, highly precise system. So far, the fixed robots used in industry are either used through a control stick or keyboard while mobile robotics mostly are autonomous. Many warehouses, such as Amazon and Walmart, have already employed mobile robots to manage the storage and retrieval of items for faster and efficient deliveries of packages and delivery services. Though these robots are not all-powerful and still need humans to work with them, having a multi-modal interactive capability may enhance the operational efficiency of these and similar industrial applications. While individual modalities may be used in case a user prefers one way of interaction than the other, traditionally, data fusion is employed to create a multimodal system through the fusion of several modalities by allowing the system to decide the accuracy of modality and combined them for interaction. Multimodal data fusion (MMDF) integrates the input of different modalities to enhance the strengths and reduce the deficiencies of the individual inputs. MMDF engines are used to perform these integrations and are used to interpret the various data streams which have different applications in different scenarios depending on the user, time, context, and task [3]. MMDF can be performed at the sensor, feature, decision or matching-score level. Sensor fusion is the combination of input streams coming directly from sensors such as webcams or microphones to form a composite input. In feature-level fusion, the features of individual modalities are combined before the decision is made while decision-level fusion involves combination after the individual decision of each modality is available, and score-level fusion generates a mean score of all modalities based on the individual scores of each input before making a decision. In the twenty-first century, multimodality is gaining a lot of attention owing to its advantages over unimodality, such as making use of more senses that connects machines more naturally to the humans, allowing new ways of robust interaction that are fast and efficient and allow respective disambiguation of identification errors [4]. The primary motive is to enable systems to work either standalone or fused with other modalities depending upon user accessibility. Multimodal interaction is part of everyday human life; we talk, use gestures, move around, and shift our gaze for an effective flow of communication. A good example is the use of Google Maps application while driving where the user cannot type the address but can input using speech. Moreover, the application is robust enough to accept input through both modalities [5]. The research done in these areas requires very robust and precise system architecture and hardware, which should remain relevant for several years. Although multimodal systems are supposed to be user-friendly, newly developed systems might require user training to allow the users to become proficient. The interaction with any system requires a significant cognitive load to understand the system, and there is a risk of making errors that can break the system at any point. The system's robustness could be increased by employing several mechanisms, including but not limited to training, fault tolerance, data storage scalability, sophisticated UI, logging, and learner mode for new users.

This work is a comparative study along with experimental setup its primary motive is to analyze the state-of-the-art progress in the field and note its accomplishments & gaps to extend and implement it in broader domains. Moreover, the survey also attempts to explore multimodal applications categorized by a combination of modalities used. The research concludes with the analysis of multimodal systems and their pros and cons in terms of reliability, usage, and performance, along with an experiment to fuse input modalities for a robotic arm. In the experimental setup, we have created a prototype robotic arm capable of handling input from different modalities and perform actions as per user's requirement.

1.1 Problem Statement

One might ask: what is HCI, and how does it work? Human interaction involves the use of technology to form an interaction with humans and a technological device. Humans can interact with computing devices in many ways while working on a dissertation, as a user is interacting with a computer while using the keyboard. The keyboard is one of many devices that are being used worldwide to interact with any computer. One of the most wonderful and least appreciated things about technology is the diversity of devices. However, in the last few years, technology has evolved and added modalities such as touch, gestures, or even voice interactions. Having multiple devices that create this interaction has changed our world and they ways we use and interact with all this modern technology. Both humans and computers play a beneficial role for each other by creating this interaction; multiple tasks can be carried out. Although it all sounds perfect at the moment, HCI devices have both pros and cons. The Human-Computer Interaction requires constant improvement to function properly according to the users needs. So essentially, the improvement of usability and how the usability should be understood as per the user's requirement, how it relates to other social and cultural values, and considering the fact when it is required, and when it may not be a desired feature of interfaces, all these factors should be discussed in detail before designing any product. The research goal is to improve the ways robots used in industry so that a person can stand far away from an actual site and give commands using gestures and speech. It will not only make life safer for a person operating a robot but also provide him or her the flexibility to use various input methods.

1.2 Motivation

In our everyday lives, we interact not only with people but also with devices such as phones, computers, tablets, and even cars. As technology advances, it become an inevitable part of our lives. HCI is an interaction between a user and a computer/device in which they work together to accomplish certain goals. These interactions are used not only for business purposes but for also downtime, social networking, communicating, and much more. Humans are constantly interacting with gadgets, especially cell phones. Today's technology is advancing at such a fast pace that we are not only influencing it, but it is also influencing us. We see everyone on the phone, using a computer, or interacting with some tool that involves technology. This is especially true when it comes to the different sensors that are used when interacting with them, such as touch, voice, motion, and hearing. These sensory products are made for usability and functionality to create a better HCI. These input plays an important role while developing a new product. Factors such as the purpose and Graphical User Interface is hugely important when developing a product. The motive behind creating new products is to make interaction easier as well as more enjoyable, effective, and useful. The influence that technology has on us is very impactful and meaningful that it contributes to the way we act and feel. These modern advancements of HCI are the reason we are going to build a product using the technology that is out today and will show how we can improve to make human life better. The research is in line with the motto of the University of Toledo, which states, "Every Rocket has A Mission, to Improve the Human Condition."

1.3 Objectives and Contributions

The objective of this research is to evaluate the strengths and weaknesses of the approach taken for various unimodal and multimodal devices available in the market. Further, we explore how we can control the robotic arm by fusing the input from several modalities. The contribution of our work is manifold, as summarized below.

- Conduct a literature survey of multimodal and unimodal systems by analyzing the pros and cons of their design and usage.
- Describe the features a multimodal system should possess to achieve a goal.
- Elaborate the experimental details and results achieved by combining speech and gesture input to a robotic arm.
- The possible application areas where our research could be put into use are also discussed.

- Publications as shown in Table 1.1

Table 1.1: Contribution

Type	**Contribution**
Grant Submitted-NSF	'Mentor Fueled Computational Thinking' with an Estimated budget of $299,052
Journal Paper	Applications of Multi-Modal Data Fusion in Enhancing Human-Machine Interaction: A Survey.(Under review)
Journal Paper	Decision-based Multimodal Data Fusion for controlling Robotic Arm using EMG and Speech. (Under review)
Journal Paper	Incorporate Computational thinking in High School Curriculum. (Under preparation)
Conference Paper	Multi-modal data fusion of Voice and EMG data for Robotic Control [6]
Conference Paper	A Real-World Implementation of SQL Injection Attack Using Open Source Tools for Enhanced Cybersecurity Learning [7]
Conference Paper	Remote Desktop Backdoor Implementation with Reverse TCP Payload using Open Source Tools for Instructional Use [8]
Conference Paper	Simulation and Analysis of DDoS Attack on Connected Autonomous Vehicular Network using OMNET++

1.4 Dissertation Outline

The thesis unfolds as follows:

Chapter 1 introduces the dissertation topic, the problem statement, and the motivation behind it. It also discusses HCI briefly and how it is relevant for our research work and, finally, describes the research objective and the major contributions of this work.

Chapter 2 discusses unimodal & multimodal systems with and without fusion in detail and surveys the state-of-the-art applications implementing combination of modalities, including pros and cons

Chapter 3 discusses the experimental details of controlling the robotic arm using fusion of speech and gesture.

Chapter 4 presents machine learning implementation of training the robotic arm using speech and EMG data.

Chapter 5 concludes the dissertation by summarizing major results and findings obtained in this research and gives recommendations for future work. It discusses possible extensions of our developed application.

Finally, the thesis ends with Appendix A, containing the C# source code written to design and develop the interaction with the robotic arm..

Chapter 2

Literature Survey - Unimodal & Multimodal Systems

2.1 Unimodal Systems

Unimodal systems are designed with single-channel input and are thus confined to a single mode of HMI [9]. Three popular broad categories of unimodal systems are based on haptic, gesture, and speech. Common examples of interfaces used in such systems are textual, graphical/video and touch. Unimodal systems implemented in the 1990s were widely used in the automation and health care industries. The unimodal systems used in surgery during the early 1990s [10] and are capable enough to understand pre-recorded voice commands and accept only single modality, i.e., speech. They possess many similar limitations; among them a prominent one is accepting pre-defined 'canned' inputs and severely lack in accepting dynamic human speech. Some example robots included MAIA [10], RHINO, and AESOP.

In most cases, the human has to initiate the dialog; the systems do not support the flexibly mixed initiative. The robot is incapable of locating the robot physically and unable to respond with their location coordinates; in a few cases, the robots support a canned feature. Robots are unable to handle effective speech; that is emotions are neither perceived nor produced. These robots are only capable of handling speech with only a few pre-defined commands. Their non-verbal communication abilities do not exist; for instance, gestures, gait, facial expressions, and head nods are neither perceived nor created. The robots do not support machine learning; they do not learn from the data provided or generated by them. [10].

Unimodal systems are broadly classified into four general categories, which are haptic, gesture, visual and speech. In our analysis, we investigate each of them. The conceivable strategies utilized for visual are face location, gaze, facial expression, lip-reading, face-based identity, and other client attributes, for example, age, sex, race, and so forth, while voice is actualized through speech input. The other input strategies, for example, haptic and gesture are accomplished through pressure, touch, and nonverbal communication.

2.1.1 Haptic Feedback

Haptic feedback is a field of research exploring human perception and interactions facilitated via the sense of touch, comprising hardware and software able to deliver touch feedback. Haptic communication refers to the use of artificially formed haptic prompts as a medium for communication between two or more individuals [11]. Multi-touch devices have established acceptability in public spaces, with huge displays appearing in markets, educational institutes, commercial residences, and other areas of high traffic concentration. These systems are designed to adopt changes over time on their interface, allowing users to interact with minimal input from the user [12].

This sense of touch via haptic feedback finds application in a variety of consumer handheld devices which are in turn used by a variety of people. The use of haptic feedback and its effectiveness with one particular group, older adults, was explored in 2015 by ECOMODE in Trento, Italy [13]. The study explored how elderly people interact with portable devices along with how they utilize various applications based on mid-air input interaction. The fourth generation of hand-held devices, which comprises cell phones, tablets, and PDAs, heavily relies on touch input. Various input methods are used by these devices which includes pinching, swiping and double click which is not familiar with older adults. In the initial survey performed by ECOMODE team [13] on elderly people in order to get their feedback in learning new technologies and challenges faced by them while using mobile devices, the researchers found that the subjects were not always disinterested in the technology, but they were very interested in using the technology to interact with relatives or to get useful information. The challenges faced by the researchers when dealing with the elderly subjects included unfamiliarity with charging ports and fragile ON/OFF buttons. There were also dexterity issues while attempting to use touch gestures with small icons, and users with low vision also made the hand-held devices less user-friendly.

An experiment was performed using ECOMODE on six elderly people in which they were asked to click on ten photographs using a Samsung Galaxy S5, an iPad mini (eight-inch tablet), and a Samsung Galaxy Tab S 10.5 (ten-inch tablet) [13]. The feedback received indicated that most users preferred the larger screens of the tablet over the smartphone. The other issues raised by the experimental subjects as follows:

1. Lack of clear feedback after clicking the photograph
2. Presence of ambiguous items on the desktop
3. Presence of extra cover/stand on the device
4. Difficulties due to reflections on the screen

There is no empirical data provided by the authors; rather, it is simply a proposal that explores how to make handheld devices more usable for elderly people.

In an autonomous vehicle, for better control inside a car, haptic features Figure 2-1 are being added to enhance the feedback a person gets in a vehicle. Bosch showcased gesture control to help control different functions in the car. Ultrasound waves that hit the hand, making it feel as though there is a knob there, but there is no physical knob [14].

Figure 2-1: Haptic feedback

2.1.2 Gesture Input

Gestures are expressive and meaningful body motions involving physical movements of the fingers, hands, arms, head, face, or body with the intent of transferring significant information while communicating with the system [15]. Research has been carried out to improve the accuracy of the gesture input. The researchers have used, a Baxter Research Robot (Yogi) and a PR2 (Kodiak) [16]. The research deals with improving grasping capability of a robot. In this experiment, they have used Microsoft Kinect on top of Baxter robot angled downwards at roughly 75 degrees toward a table in front of it, which enables the robot to detect things and pick them up from one point to drop them at another. The empirical data shows the results are 84 and 89 percent accurate with Baxter and PR2, respectively. Gestures were used in another implementation called ModDrop, which is based on adaptive gesture recognition. The term ”multi-modal” over here refers to various gestures captured through the left and right arm both arms are treated as different modalities. The research focuses on gesture-based detection on multi-scale and multimodal deep learning [17]. The researchers have captured spatial information on users to initialize the modalities carefully and to fuse them for cross-modality connections while preserving the uniqueness of every modality.

Ferron et al., (2015) while working with older adults using ECOMODE, finds cell phones can also use mid-air gestures for interaction, interestingly it is much appreciated by them expressing positive comments the possibility of using mid-air gesture [13]. MYO armband were provided to two elderly ladies, and both of them found this interaction modality interesting. They can control a music application using a MYO armband. Other experiments have been performed that also leveraged gesture input. In one such experiment, the developers attempted to build a healthy relationship between a robot and a human being. Like most of the researchers within the human-robot interaction field, they also utilized the Microsoft Kinect. The initial application for the project is in the homes of seniors where humans arent easily available to help the resident. Robots would be deployed to improve the services provided to them.

The result of the research project was a service robot called Donaxi. The robot has an omnidirectional navigation system [18] with four wheels (each one containing a DC motor and encoder) and a laser system on the front and back for mapping and navigation. The robot is equipped for understanding both voice and motion utilizing the Microsoft Kinect, yet they have utilized motions for the implementation. The development team has participated in the Mexican Robotics Tournament (TMR2015)

[18]. The Donaxi robot is trained with many videos from different people in different places, which enables the system to work with a variety of users and environments. After providing gesture input to the Microsoft Kinect, the system was trained completely. The above mentioned 2.1 explains that it requires enormous effort to train the system, that is attention is needed for 417 iterations to get trained properly, which is a substantial effort for a single user command. The experimental results show that once the system is functioning, it would able to recognize Attention, Stop and Right for any user as shown in Figure 2-2. The work mentioned is not complete, and more features are planned, including the understanding of the additional gestures.

Table 2.1: Total Number of Gestures in the Dataset

Gesture	Stop	Come	Left	Right	Attention	Indication	Turn
Number	194	177	395	407	417	363	207

(a) Attention (b) Stop (c) Right

Figure 2-2: Three Gestures Used at TMR2015 for Donaxi

2.1.2.1 Leap Motion

Leap Motion is a motion control device that connects to a computer and enables users to manipulate objects with their hand motions. The programs are designed to recognize and interpret gesture-based computing to create designs, play games, or carry out some other type of task. Leap Motion is a real-time interaction that can manipulate digital objects. You can run Leap Motion on devices like MAC and Windows. This virtual reality is a great tool of the future, but it is not nearly where gesture control needs to be for the HCI to be smooth. A new update called Orion has provided some recent updates to the fourth-generation core software to improve the finger tracking and motions, faster and more consistent hand initialization, and more accurate shape and scale for the hands just to name a few. The hardware is made with two cameras and three infrared LEDs, which track light with a wavelength of 850 nanometers. The interaction areas are eight feet apart while the motion controllers are 2.6 feet apart after the Orion update. Having gestures like these makes it difficult to interpret or transcribe items like bulbs, daylight, and halogens, which would light up the scene [19]. The Figure 2-3 shows how the sensory object orientation of the VR works.

The Leap Motion controller is a small USB device that plugs into your computer. Using LED lights and camera sensors, the Leap Motion controller scans an area of eight-cubic feet above the device. It tracks both hands and all ten fingers as they move through the open space between you and your computer the special software senses your hands and fingers and translates the data into information for your computer. Leap Motion was developed in 2008; the Leap Motion controller is a small USB peripheral device designed to be placed on a physical desktop, facing upward. It can also be straddling onto a virtual reality headset. Using two shaded IR cameras and three infrared LEDs, the device observes an unevenly curved area, to a distance of one meter. This is then sent through a USB cable to the host computer, where the Leap Motion software analyzes it. Leap Motion initially distributed thousands of units to developers who are interested in creating applications for the device. The Leap Motion controller was first shipped in July 2013. In February 2016, Leap Motion released a major beta update to its core software. Dubbed Orion, the software is designed for hand tracking in virtual reality [20].

Figure 2-3: Leap Motion

2.1.2.2 CaptoGlove

This glove is a haptic interface system that is Windows compatible and also works with iOS and Android apps. Haptic technology is any which can create a sense of touch by applying force, vibrations, or motions to the user. This Capto Glove Figure 2-4 is a motion controller that works using Bluetooth technology. The CaptoGlove claims to be compatible with most VR headset that is already on the market. It is rechargeable, and the battery lasts ten hours. The glove has movement sensors in each finger and a pressure sensor on the thumb. It can be used to play VR, PC, and Phone games or as a controller for many devices and many platforms. You can use just one glove to control and interact, or you can buy both gloves, they cost $250.00. It claims to be able to control any past, present, or future game created. Most haptic devices are made to interact with virtual reality environments and have sensors that allow you to control and give lifelike feedback. For example, you could probably use the glove to control a car in a racing game, and if you hit a wall or something, it will probably vibrate or shake.

Figure 2-4: Capto Glove

2.1.2.3 MYO Armband

The MYO armband is a wearable gesture and motion control device that lets you take control of your phones, computers, and other devices touch-free. Electromyography (EMG) is a technique for evaluating and recording the electrical activity produced by skeletal muscles. The MYO armband lets you use the electrical activity in your muscles to wirelessly control your computer, phone, and other favorite digital technologies. A simple wave of your hand will transform how you interact with your digital world [21].

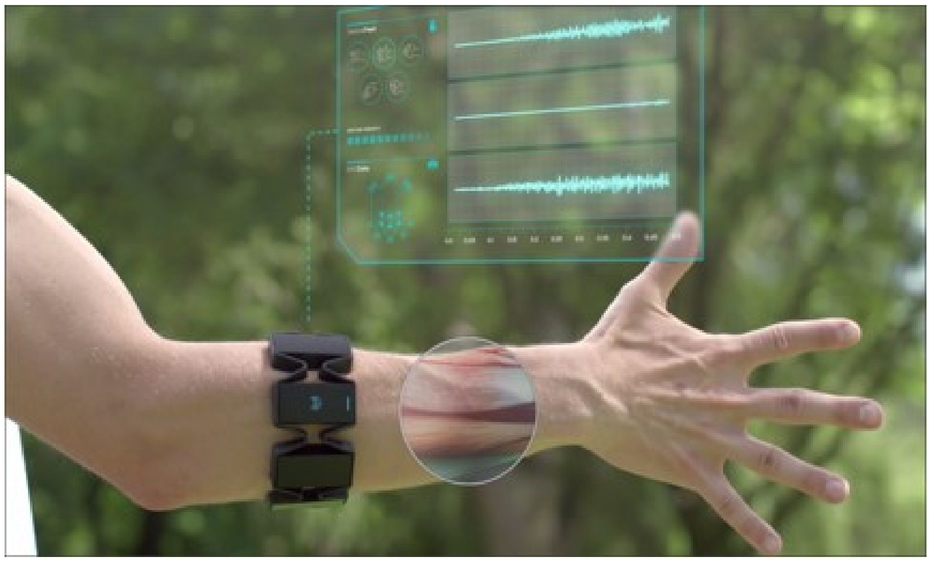

Figure 2-5: MYO Arm band

2.1.3 Speech Input

The last two decades have seen the development of increasing possibilities where computers and handheld devices smart devices can be used [22]. Among them, the feature which was introduced off lately is speech recognition apart from trivial use of keyboard and mouse. Speech recognition is an analysis of the human voice for performing a certain task on a device. Pursuing a similar concept, a team from MIT is working on enabling human-robot interaction. In this research, a team from MIT and Germany worked together and developed a system capable of understanding and adapting human speech by a robot implemented in a wheelchair [23]. The system is efficient as it starts the process again whenever it finds any of the following input issues

1. User utterance is inconsistent with current discourse (unification with discourse info fails).
2. User utterance can only partly be parsed.
3. User utterance is inconsistent with the robots expectations (unexpected info).
4. User asks for the same info several times.
5. No speech can be found in the user utterance.

The system consists of a speech recognizer, a natural language parser, and a dialogue manager. The drawback of this system is that it only accepts speech as an input, while they have claimed it as multimodal [23].

A newer feature, created a little less than ten years ago, that has greatly changed our interaction with computers is the voice assistant. Depending on the device or OS, assistants like Siri, Cortana, and Alexa have made interaction with our devices a little easier. They recognize your voice and can be summoned with a simple phrase. The voice recognition software can understand numerous commands and can help you carry them out hands-free. It can start many apps, schedule appointments, return calls or texts, search the internet for the user, and can also give you directions and much more through voice commands. Alexa is set up in your house, and she can play songs for you on demand. She can also place an order through Amazon Prime for you. If you have lights or appliances set up on smart switches, you can ask her to turn them on or off for you. Speech recognition can also help you drive safer because you wont have to text and drive; you can prompt Google or Siri to compose and send a text to one of your contacts [24].

2.1.3.1 Siri

Siri is a technology that uses voice commands. In everyday life, Siri is used across the world on Apple products. An iPhone user can use Siri for just about anything, as long as your phone is set up to recognize when you say "Hey Siri". When the command is said, the voice recognition program responds with, "What can I help with" as shown in Figure Figure 2-6 The program can also be given the command as soon as you say, Siri, simple commands like play a song, call this person, put this meeting on my schedule, and can even be asked sophisticated questions that will direct you to a site that can help you. Siri is one of the most used virtual assistants out there today, with different languages you can set it too and different accents as well [25].

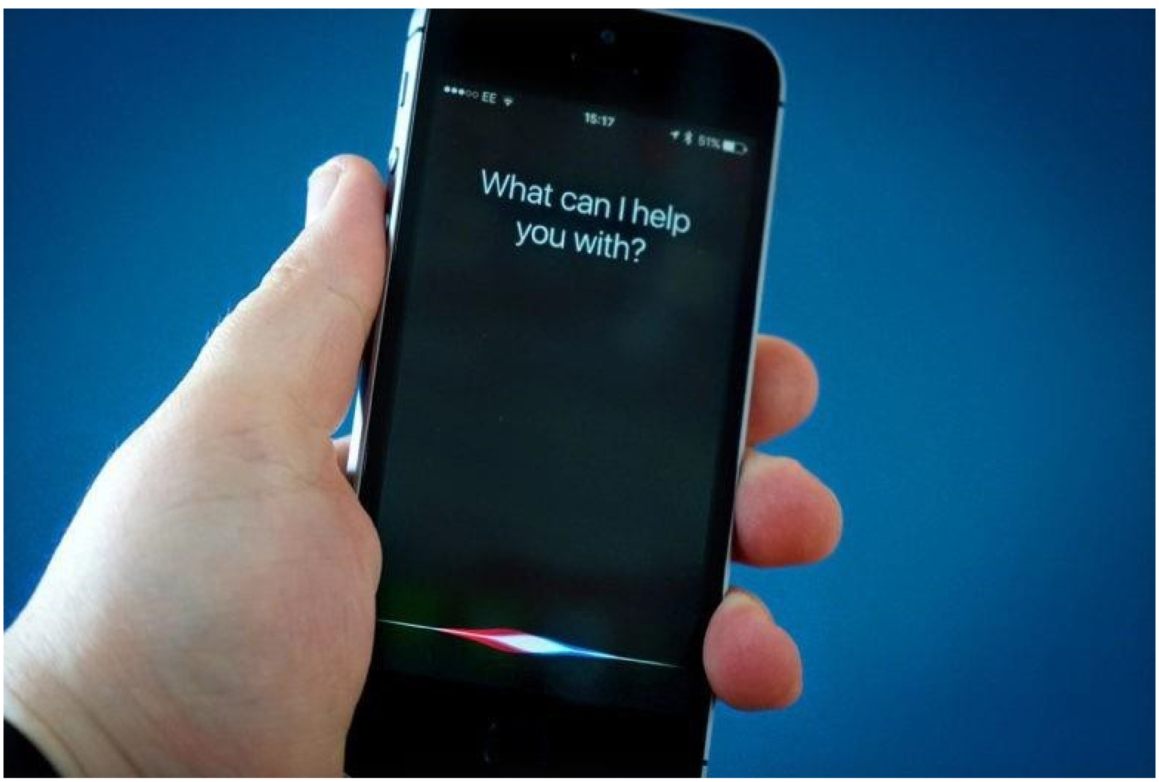

Figure 2-6: Voice recognition using Siri

2.1.3.2 Cortana

Cortana is Microsoft's virtual assistant. Cortana is activated by voice commands and is capable of performing several tasks on your hand-held device. Cortana's biggest market competitor is Apple's Siri. Most users find Cortana very useful for gaming consoles and recommend that everyone take advantage of this helpful tool. Cortana can access the internet and do a lot of things that you tell it to. In the past, the users can easily command Cortana to record the gameplay and it is quite easier to tell Cortana to record that instead of going to the settings and clicking options, then record, and so on. One cool thing about Cortana is that you do not need a Kinect to make it work, and the player can use a microphone connected to the Xbox controller to make it work 2-7. Computers also use Cortana, and it is very handy when looking for something on your desktop or even search the web. Cortana is a huge advantage for people that have problems with typing or maybe even some sort of disability. It is still used a lot; though less common in cell phones than video game consoles. The review suggest, more people discusses Siri and use it instead of Cortana. Still, Cortana is used today in several devices, and the research suggest it is doing well because of substantial user base for Microsoft Windows and youngsters who play games using Microsoft Kinect or XBox. Cortana is also free with an Xbox live membership, but prices vary. One year is $60, one month is $10, and three months cost $25 [26].

2.1.3.3 Amazon Echo

Another popular tool that individuals have been using since 2014 is Amazon Echo. Amazon Echo is a device that is capable of being your personal assistant. It provides information data from the world wide web in real time. It is a conversation voice-control tool that could be used to ask questions, play music, and control other technology devices such as lamps and speakers. Amazon Echo could be integrated into your smart home devices to control the temperature, lock the doors, and also

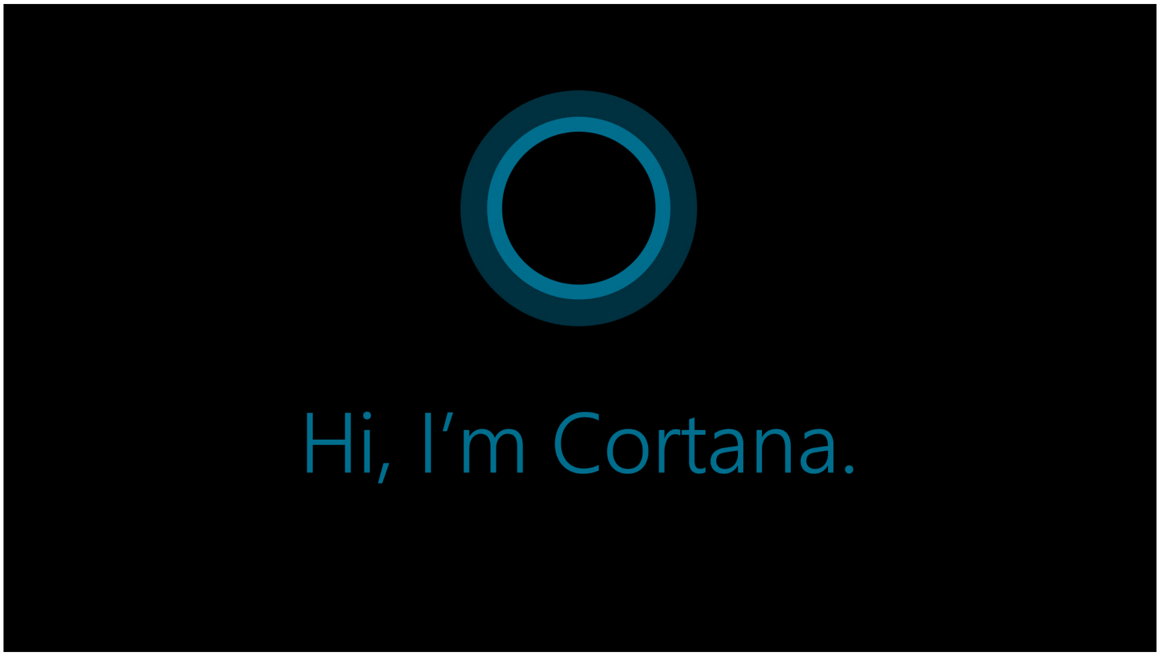

Figure 2-7: Cortana

dim the lights. Much like the iPhone tool, called Siri, it can carry on conversations with a person and is constantly ready for someone to talk. To activate it, one must say "Hey Echo" or "Hey Alexa," which will enable the assistant to give voice commands to perform a specific task. The device will start to listen to the area nearby for some type of response to carry out the task. This voice assistant control is known as AVS (Amazon Voice Service), which is an intelligent voice-recognition device that can understand humans. Some recent Echo devices such as the Amazon Look have a built-in camera to take pictures and videos. For example, the sensor is made to learn one's taste by taking photos of different clothes one wears and to have them a better shopping experience using machine learning techniques [26]. The Figure 2-8 shows the layout, and sensory location of the Amazon Echo Look.

Figure 2-8: Amazon Echo

2.1.4 Eye Gaze Input

The Eyegaze Edge, created by LC Technologies, is a device that gives users who are not able to use their hands a way to communicate with computer just with the movement of their eyes. The way the technology works is by calibrating the irises with the screen. A small calibration point moves around the screen as the user follows it with his or her eyes and then the eye is entered into the system; this is a quick and easy process. There is a low light, an infrared camera that focuses on the eye and takes you through motions to get a good reading of your eye movement. The Eyegaze Edge Figure 2-9 is built so that the user can calibrate with one eye and navigate with one eye as well. Different means of communication can be used on the system as well including picture icon boards, prestored phrases or store new phrases, computer keyboards for emailing texting or simply taking notes, and even connects to your computer or phone. All this is possible through the image processing software that can determine throgh analysis, where the user's eyes are going to move next [27].

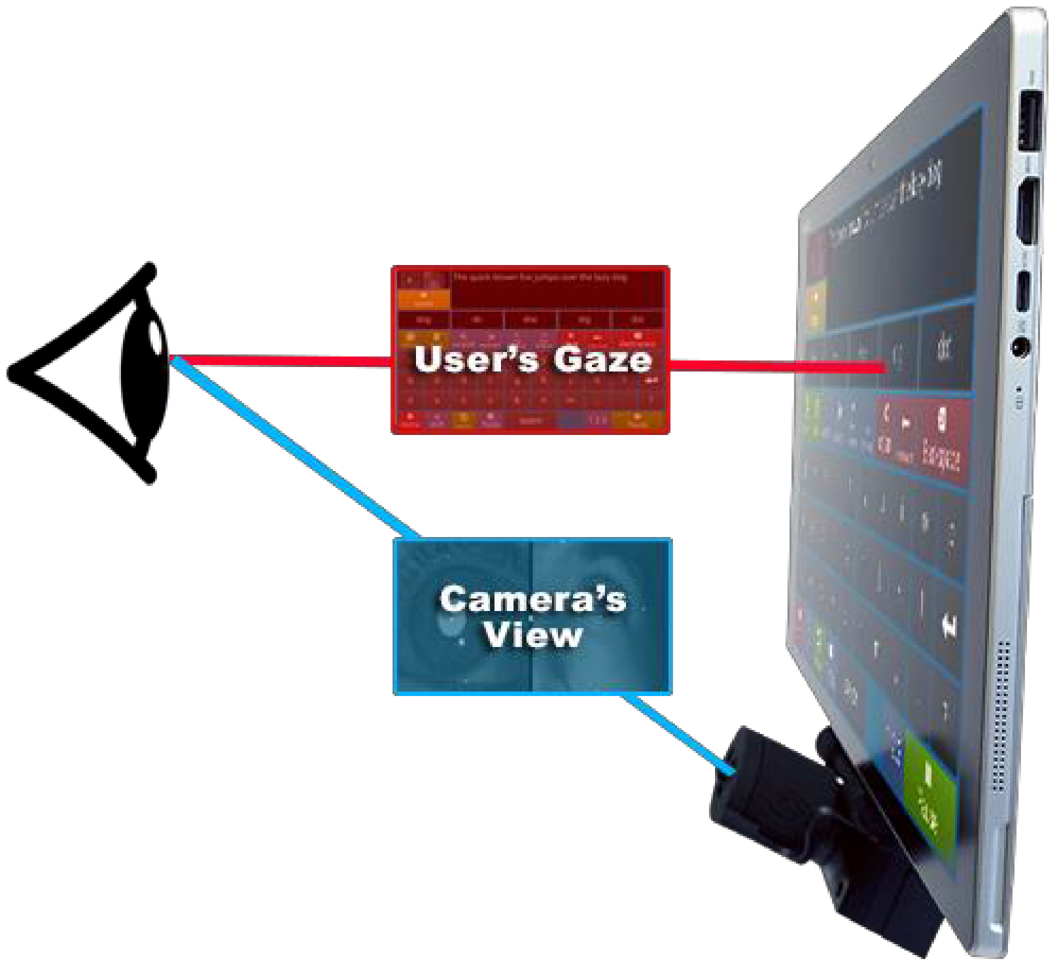

Figure 2-9: Eye Gaze Input

2.1.4.1 Tobii EyeTracker 4C

Eye-tracking technology is a rapidly growing area for both gaming and research, and Tobii is a company that produces many popular devices in this area of technology for both fun and research. Eye tracking Figure 2-10 is the process of measuring the point of gaze or the motion of the eye, relative to the head. Eye-tracking in HCI investigates the scan path for usability purposes or as a method of input in gaze-based interfaces. Although eye-tracking devices can be used for gaming, it seems that the majority of eye-tracking devices are for mainly used for research purposes. There are many types of research being performed with this technology that seem fascinating. The data from eye tracking devices are being used in psychology and medical research; the data gathered is also being used in marketing research. This particular product is a camera that is $150.00, and it can perform several tasks, but it also depends on what software you purchase, and they compatible with your device. This camera streams and shows your viewers exactly what you are looking at in real time with ghost software. It is compatible with games available through the same company; it supports more than 150 games. It also has a training tool that is supposed to improve your game-playing skills. It uses a USB connection. It is also a way to securely login to Windows 10. It has some high ratings and good reviews. The main complaints were that it was hard to set up and had hiccups in the games [28].

Figure 2-10: Tobii 4C Eye Tracker

2.1.5 Touch Input

One of the most common modalities we use daily to interact with computers is the touch screen. The majority of people of all ages know how to use the touch screen. After the introduction of the smartphone, along came tablets. Once they became common in most households, then the touch screens became one of the most popular ways to interact with computers. When you use self-checkout at the store, get money from the ATM, place an order at McDonalds, or rent a Red Box movie, you need to know how to use a touchscreen to complete these tasks. Whether a user likes it or not, CRT and LCD screens are everywhere and have changed the way we interact with computers. Touch screens have given us the ability to use various applications; they are not only used in public information systems such as kiosks at one of the local restaurants or ticketing machines when the user visits the Motor Vehicle Division (MVD) office or any other offices; it is also on devices such as iPhones, tablets, touchscreen TVs, or even touch screen refrigerators, dryer, and washer. As we can observe, the devices are pretty much all around wherever we go, whether it be to wash our cars, to get a parking ticket and find any restaurant; the screen of our PC or even the screen on our vehicles that used for GPS, calling and entertainment purposes or so forth. It is scary to think that our touchscreens are replacing our conventional buttons because of the way they operate. Think about when we had our flip phones and all the buttons that were on them, and after decades, all those buttons have become rare because they are being replaced with the marvelous invention called touchscreens.

There are different types of touch screens; they differ in hardware and software. The ones we encounter when use kiosks at the stores and ATMs use resistive technology Figure 2-11. These screens have two thin layers, one that is resistive and one that is conductive. The screens have a gap in between them with a constant electrical current running through that gap. When you touch the screen, the two screens touch, changing the electrical current, the software read the current change and carries out the instruction related to those coordinates. There are also capacitive touch screens Figure 2-13, which are made of materials that hold an electric charge in wires thinner than a hair, arranged in a grid. There are two types of capacitive touch screens; surface capacitive and projected capacitive. They both work similarly; the main difference is that a projective screen has a separate chip for sensing. They work by transferring an electrical charge to your finger, when you touch the screen a circuit is completed, and a voltage drop occurs at that location on the screen. The software then carries out whatever task is related to the location of the voltage drop. Touch screens have revolutionized the way we interact with computers, making it so those of all ages can interact with a computer and changed the way we complete daily tasks and errands. There are several devices with finger touch ID (capacitive fingerprint scanner), like computers and phones [29]. Though a user would think that finger touch ID would use light technology, in reality, the capacitors use electric currents from the spacing the in ridges of your fingers as shown in Figure 2-11. The electricity sends a pulse and gets your print. When the correct fingerprint is read, it will unlock the device, making it much harder for someone to hack you or break into your computer [30].

2.1.5.1 Samsung Foldable Phone

The Samsung foldable phone is a recent invention that is not yet in the market. It is a touch screen-based devices which accepts input on both sides. The phone was introduced early in 2019 as the first real functioning foldable phone. This dual-battery phone is the first of its kind that brings back the foldable slick aspect of older-generation phones. The phone is one of the most expensive phones in recent times, costing $1,980 to $2,600. The Galaxy Fold is made with two foldable screens.

Figure 2-11: Touch Input

The display is made with an ultrathin polymer that uses a new adhesive made by Samsung which enables it to fold many times called an Infinity Flex Display [31]. The figure shows the entire device with the two screens capable of folding Figure 2-14.

2.1.6 Game Controller Input

Video games play an integral role in youngsters, lives these days. They spend a daily couple of hours each day on a network-based gaming environment with their classmates and friends, playing games from a remote location.

2.1.6.1 DualShock 4

DualShock 4 Figure 2-15 is a new hardware that supports a PlayStation 4 controller. It helps the user to navigate through the system and interact with the game and even online with other users. This control has many different features like a

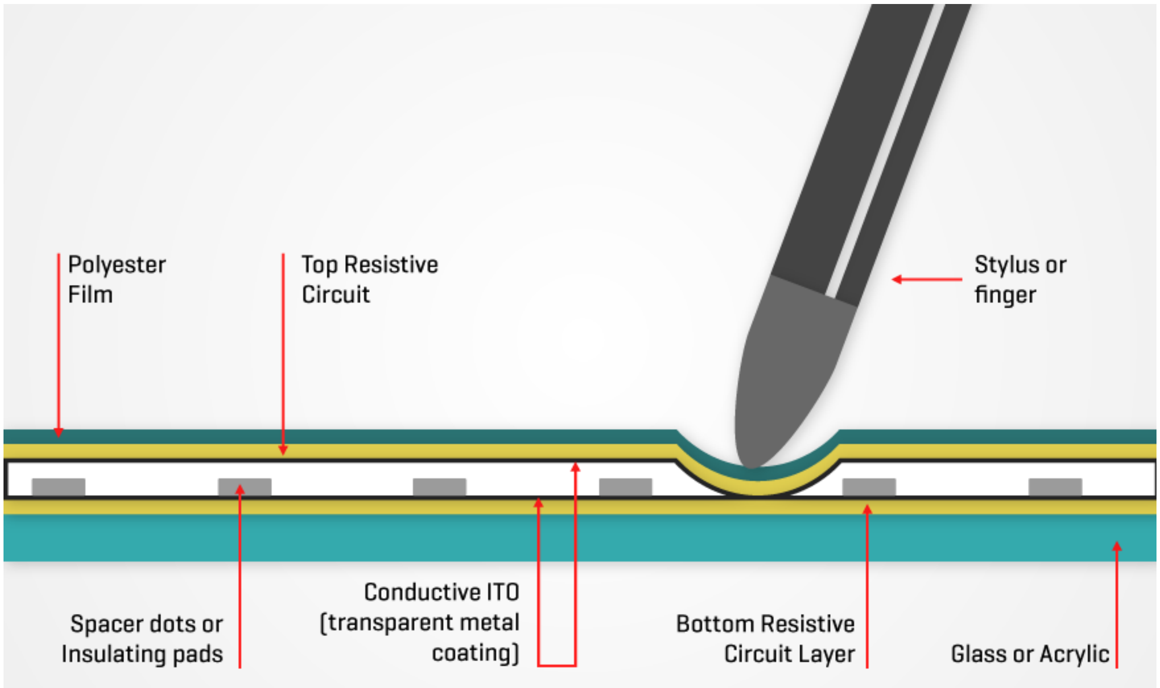

Figure 2-12: Resistive Touch

touchpad that makes it easy to navigate on the internet like a computer touchpad Figure 2-15 . It has vibrations that help players in a game know they are in danger; there is a light bar in the front indicating that the controller is on. This controller is also wireless and rechargeable, just like a cellphone.

2.1.6.2 Nintendo Switch

Nintendo Switch as shown in Figure 2-16 is a gaming console that was released in 2017, although Nintendo is the oldest market player for gaming consoles. The Nintendo Switch is a very successful console because of the classic games that the old-school Nintendo came with. You can use the switch as a remote control for your tv to turn it off and on and to switch the input on your tv. Some older TVs might not be compatible. You can find a lost controller by using another controller. The missing controller will vibrate until found. The Nintendo Switch comes with four controllers,

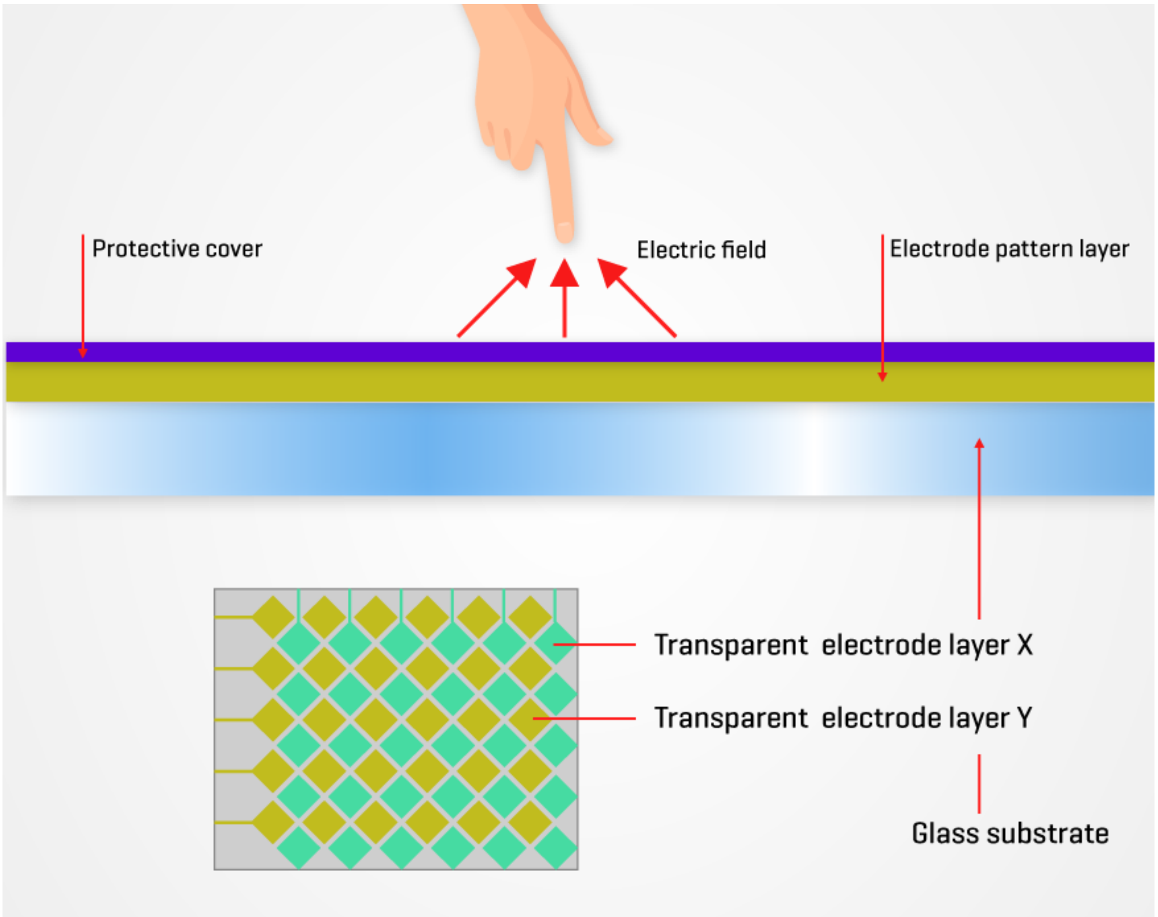

Figure 2-13: Capacitative Touch

and these controllers use different things to control them. They have buttons and triggers, but they can also be controlled by shaking them, twisting them, blowing into a sensor that detects, air and many different kinds of motions. For example, you could be playing a driving game, and you have to hold the controller like a steering wheel and make it turn, just like driving a car. The Nintendo Switch can also be a hand-held device by plugging two of the controllers into the sides of a small handheld monitor. This is the only console that you can really take on the go and also use on a big TV easily. Prices vary anywhere from $299 to about $500 for a Nintendo Switch. Gaming consoles from Nintendo are worth it because you can have a good time and the cool features. They are really cool user friendly interfaces that you can move the

Figure 2-14: Samsung Foldable Phone

controllers around and perform several inputs like you would in real life [32].

2.1.7 Digital Pen

Digital pens can transmit your writing to the computer using wireless technology. The pen is thick and packed with digital circuits; just like a mouse, it uses a photocell light detector and a LED light emitter as shown in Figure 2-17. The difference between the two is that they are stacked vertically instead of horizontally. Also, the pen keeps track of your movements and patterns. Digital pens can upload your writing onto computers though connecting a cord, connecting to a charging dock, o,r best of all, through Bluetooth or infrared [33].

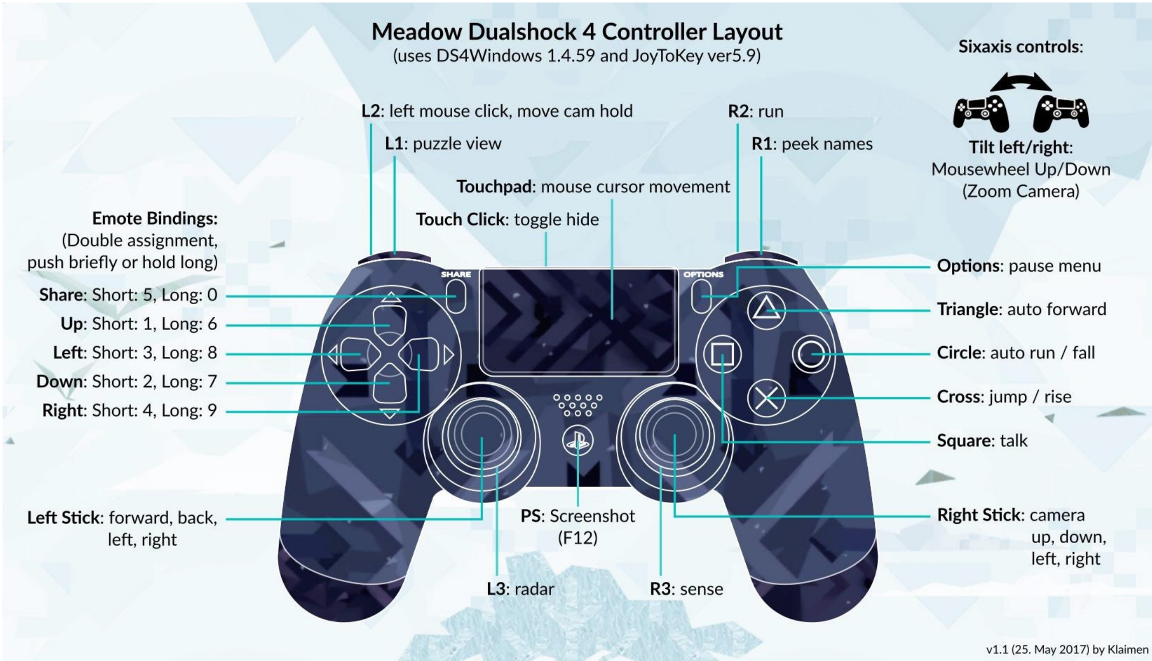

Figure 2-15: Meadow Dualshock 4 Controller Layout

2.1.8 Facial Recognition

Facial recognition is a booming technology for authentication of users from unlocking your computer to finding a criminal or even for finding a friend on Facebook, a Snapchat filter, and so on. Facial recognition takes data and stores it in its system, and when data matches or is very similar, it will recognize the face and bring up the stored face. The police and government agencies use facial recognition when searching for a person. The catch is that the person has to already have a picture in the system, or the system will have nothing to match it to; it would just compare to everyone is in the system to find the best match. Snapchat uses facial recognition to apply filters to faces and can even face swap with other individuals. The same technology is applicable for unlocking computer 2-18 and even the new cell phones.

Figure 2-16: Nintendo

2.1.9 Drones

With the recent growth of sensors and distance with network protocols, drones have become a very useful tool to use. Drones are unmanned aircraft that easily be flown remotely or fly autonomously through software-controlled flight plans working along with sensors to avoid obstacles. Drones are used for many tasks, such as search and rescue, wars, surveillance, and fun. Drones are equipped with sensors, cameras, and a software to process the data. These sensors are meant to calculate the distance, time of the flights, chemicals, stabilization, and orientation. Drones are increasingly being used in many different fields. For example, you could use a drone for journalism, disaster response, wildlife monitoring, and agricultural purposes as shown in Figure 2-19. Drones also have their downfalls as well, especially when it comes to safety and the ethics that are involved with them. drones require special permission to operate,

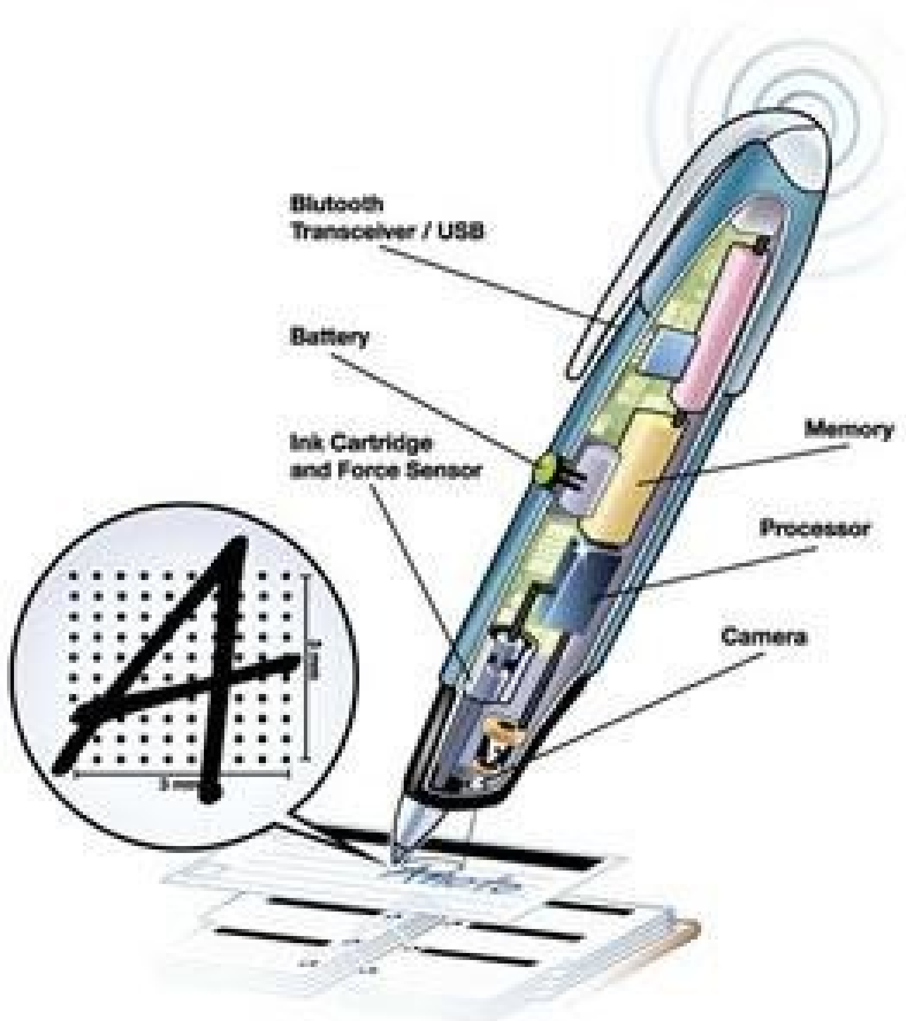

Figure 2-17: Digital Pen

and should not hamper any airspace for airlines and private jets. Recently drones have been on television and criticized for people flying their drones around airports, or also they have been seen in the Olympics. [34].

2.1.10 EPOC Emotiv headset

The EMOTIV EPOC+ is a portable, high-resolution, fourteen-channel, EEG system. It was designed to be quick and easy to fit and take measurements in practical research applications. It is compatible with all EMOTIV software products. EPOC

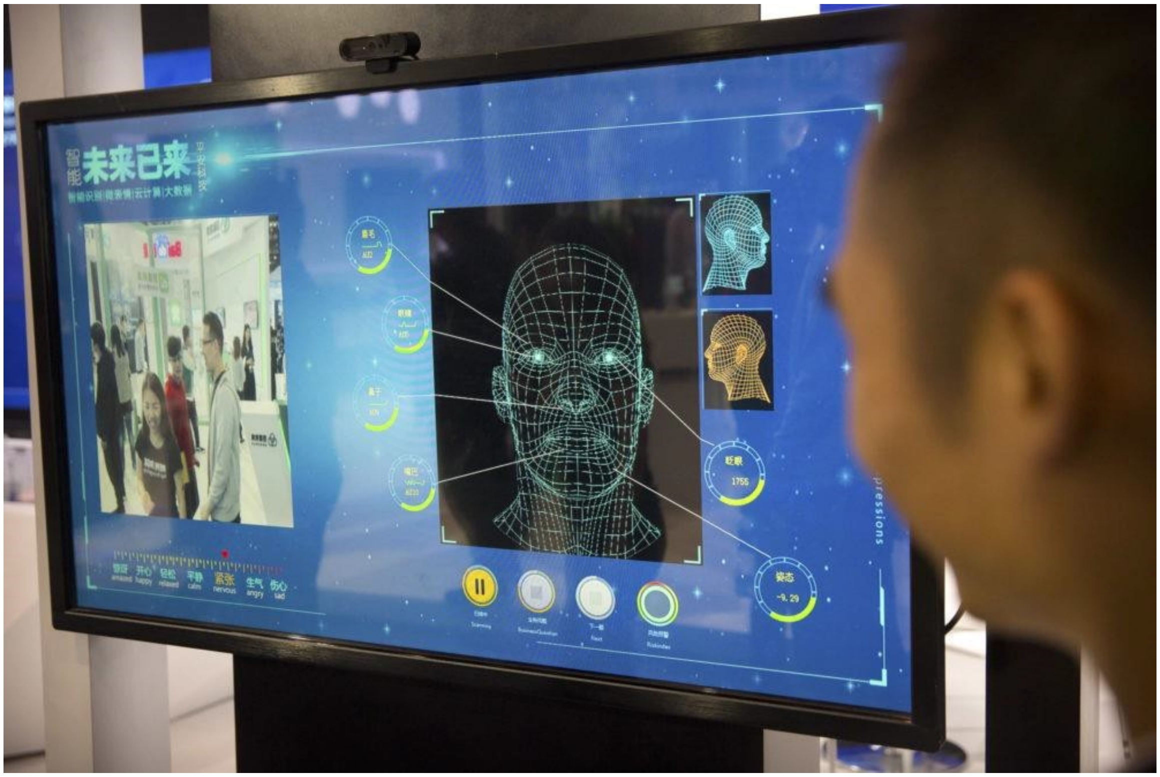

Figure 2-18: Facial Recognition

emotiv is a brain interfacing devices which accepts input through the USB. It senses the portion of the brain that is active while thinking about specific problems. [35].

2.2 Multimodal Systems

Multimodal interaction provides a user with multiple classes or modalities of interaction [36]. A simple example is defined by Bolts "Put That There" [37] which combines speech with a gesture. Multimodal systems allow the use of multiple human sensory modalities and combine many simultaneous inputs and present the data utilizing a synergistic portrayal of a wide range of output modalities [38]. Multimodality is only truly implemented when there are at least two of the following inputs in any combination: speech, text, or gesture. The usage of multimodalities such as speech,

Figure 2-19: Drones

gestures, and haptic feedback can provide a productive, and interface for a media community for various user groups (customary clients, blind users, visually disabled users, and physically impaired users). The investigation recommends the framework does not acknowledge multimodality. The input modalities can be as simple as two pointing devices, or they may include advanced perception techniques like speech recognition and machine vision. The simple cases do not require more processing power than the current graphical user interfaces, but they still provide the user with more degrees of freedom with two continuous input feeds. The multimodal inputs, in this case, are two pointing devices providing input simultaneously to a system. New perception technologies that are based on speech, vision, electrical impulses, or gestures usually require much more processing power than the current user interfaces. The review discusses recently performed research in Multimodal Systems and classifies them based on Fusion.

Figure 2-20: EPOC Emotiv

2.2.1 Multimodal Fusion

Multimodal fusion is the implementation of several modalities in an application that complements the partial input and derives meaningful results. The increasing number of multiple datasets of information, which are obtained by using different acquisition methods through peripheral devices, raised an opportunity to analyze datasets separately and fuse them to derive a common goal. However, until this decade, the data fusion technique is confined within boundaries of psychometrics and chemometrics, the communities in which they evolve. There are several surveys performed by researchers. The first that we explored was published in 2010 [39]. The analysis classified the various types of fusion: feature level, decision level, and hybrid level. Methods of multimodal fusion are also discussed, including the rule-based, classification-based, and estimation-based methods. Another study was performed in 2015 that emphasized data fusion, including various methods, challenges, and future prospects [40]. In this study, the researchers focused on multimodality in the context of multisensory systems, biomedical systems, and environmental studies. In this study, the goal is to classify the work done after 2015 differently from previous surveys, which is based on classification on the basis of the number of input methods used.

Recent advancements in technology, including a growing number of domains, leads to an increased interest in exploiting multimodal fusion efficiently [40]. Three level of fusion have been discussed, namely feature-level fusion comprising visual features, text features, audio features, motion features, and metadata. The other two levels of fusion are Decision level multimodal fusion & Hybrid multimodal fusion [39]. Decision-level fusion is based on assigning priority to each modality based on its previous experience. Hybrid multimodal fusion is a combination of both feature and decision-level strategies. The methods of multimodal fusion are classified as rule-based, classification-based, and estimation-based. The rule-based multimodal fusion is further classified into three categories:

- Linear
- Majority voting
- Custom defined rules

Linear weighted-based fusion technique is easiest, based on a first-come-and-first serve basis and linearly combining them. Classification-based multimodal fusion works on assigning linear weight to each modality, for example sum and product, MAX, MIN, AND, OR, majority voting. The classification-based multimodal fusion used the Damster-Schafer and Bayesian algorithm. For the third technique, custom-defined rules, in which the input is either customized per the input needed by the system or as per the need, there is no exact algorithm defined to accept the input. Research published in 2002 claims to implement an algorithm that fuses data of a complex system, represented as

$$x_i(t) = A_i x_i(t) + B_i u_i(t) + w_i(t)$$

$$y_i(t) = C_i x_i(t) + v_i(t), (i = 1, 2., n)$$

where: n = number of subsystems

xi(t) = state of the ith subsystem

ui(t) = control signal on the ith subsystem

yi (t) = output of the ith subsystem

wi(t) = ith subsystem noise

vi (t) = measured noise of the ith subsystem

This algorithm deals with the fusion of data received from two different input sensors [41]. The experimental results shows among the three performed experiments, only one would function while the other two would malfunctions.

The multimodal gesture recognition algorithm presents a new multimodal framework based on a multiple hypotheses fusion scheme. In this context, the multimodal input is provided by multiple users at a time. An increased number of multimodal inputs brings more challenges to this field. The challenges mentioned in this research are the detection of meaningful information in audio and visual signals, extraction of appropriate features, the building of effective classifiers, and the multimodal combination of multiple information sources. The fusion of multiple inputs can be performed at early, late, or intermediate data/feature levels. The fusion is also possible at the stage of the decision after applying independent unimodal models. This research, just like others we have reviewed before, used Microsoft Kinect, which uses color, depth and audio signals captured by the sensor. The most common approach observed in several research papers published from 2014-2016 is that they have used the Microsft Kinect for recognizing voice and gesture. We believe this is due to the technology's accuracy and precise sensors. The framework talks about accepting multiple gestures as shown in Figure 2-21 from different users and then choosing the best multi-stream hypotheses (input) as shown in Figure **??** . It involves multimodal scoring and re-sorting of hypotheses algorithm to manipulate the best gesture; once the algorithm provides the output, the gestures are performed on the system [42].

$$v_i = \sum m\epsilon S w_m v_{m,i}{}^s$$

i = 1 to L where: weights = w(m) are determined experimentally

v(m,i)ŝ = standardized version of modality scores based on Viterbi decoding.

Inputs with the highest score are selected for the next phase of the algorithm, which is called Parallel Segmental fusion. The segmental parallel fusion algorithm exploits the modality-specific time boundaries, and it observes the pattern of the sequence of input as occurred in the previous iteration. It was observed during the experiment that there is no one-to-one correspondence between segments, and they are first aligned using dynamic programming. The experimental results presented in this experiment are 93 percent accurate using Microsoft Kinect. The social signal interpretation (SSI) framework for multimodal signal processing and recognition in real-time was implemented in 2013 at the Lab for Human Centered Multimedia at Augsburg University, Germany. SSI deals with the idea of intuition in next-generation interfaces in real-time. The research seeks to enhance SSIs C++ API and provide front-end users with the ability to use text editors backed with an XML interface. Computer interfaces are based on explicit commands, but the wave of one's hand or the tone of one's voice can sometimes convey more than hundreds of such commands [43]. These natural inputs are much more capable of informing the system about the users intuition. To collect and capture human intuition requires a system that is capable of storing a collection of representative samples. Afterward, the collected data is analyzed by a skilled person who works to classify user interaction. To assemble the components mentioned above, the authors proposed a SSI toolbox. The framework proposed was part of the EU-funded Motebo project, and it dealt specifically with physiological data analysis of diabetes patients in an automotive environment using fusion. The concept is not implemented yet; hence no empirical data was provided.

A survey was performed on Multimodal Interaction at the University of California, Santa Barbara, which discusses the challenges faced while integrating various modalities into a final output and exploring the feasibility of stage, namely, whether it should be done in early versus late integration. Richard Bolt's Put That There is one of the most significant demonstrations of multimodal systems. The MIT Architecture Lab (later to become MIT Media Lab) integrated voice and gesture to present geospatial data to a user sitting in a chair. Phrases that the geospatial system understands include the following: "create a blue square there", "move that to the right of the green square" by implementing Multimodal System, "make that smaller", and the canonical "put that there". Early multimodal systems were thus focused on geospatial applications.

In 1989, another system named CUBRICON [44] was developed, which enables users to interact using spoken words, gestures, and natural language, and it displayed output in the context of map-based tactical mission planning. Later, in 1993, Koons et al. developed a system that understood speech, gesture, and gaze for a map-based application. In 1997, another system named QuickSet [44] was designed by Cohen et al., featuring pen and voice-based navigation intended for a US Marine Corps training simulator. In the post-WIMP ("windows, icons, menus, and pointer") period, the multimodal experience was further enhanced to include sketch and 3-D, which was implemented in the Butler Interface: interacting with the interface is like interacting with a human who has the ability to speak, gesture, and use facial expressions, among other forms of human communication [44], as shown in Figure

2-38. After 2000, more methods were proposed for interaction that included both verbal and non-verbal communication. Later, the concept of the Perceptual User Interface (PUI) was introduced at a workshop, which eventually grew to be one of the branches of the ACM conference. Input and output modalities, which are relevant in multimodal interaction and fusion, are further classified into modes and channels, as shown in Table 8. Their respective contexts and discrete requirements have distinguished multisensory and multimodal devices.

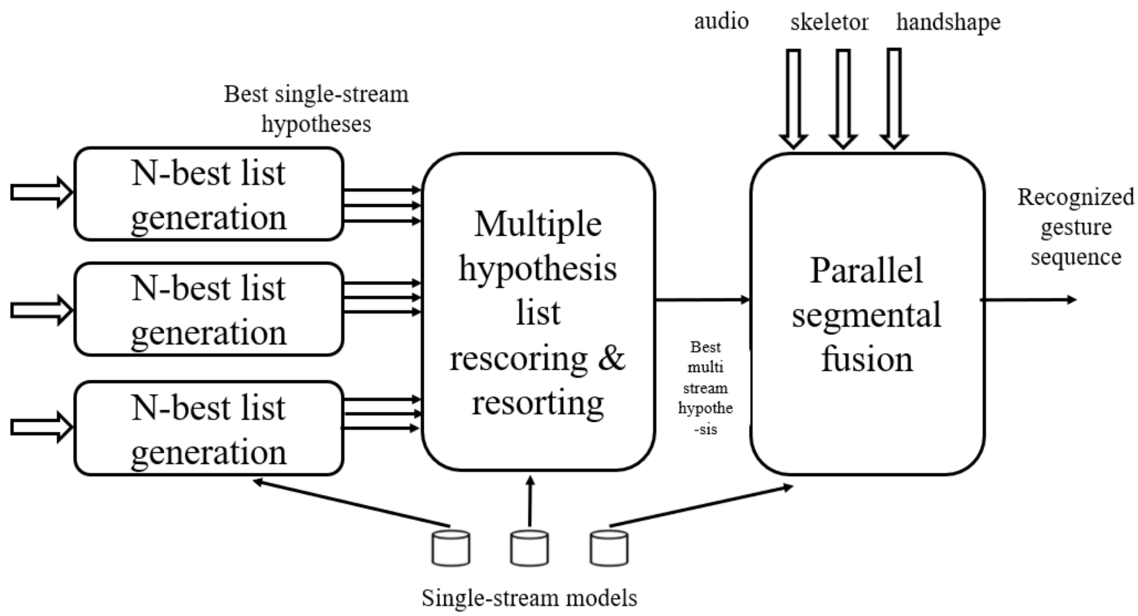

Figure 2-21: Overview of the Proposed Multimodal Fusion Scheme for Gesture Recognition Based

2.2.2 Multimodal Systems with Fusion

Speech is a rich channel for human-to-human communication and possibilities to be a rich channel for human-to-machine or computer correspondence. Gestures supplement our speech in various ways, adding redundancy, emphasis, humor, and description, and depiction. Multimodal interfaces made from speech and gesture

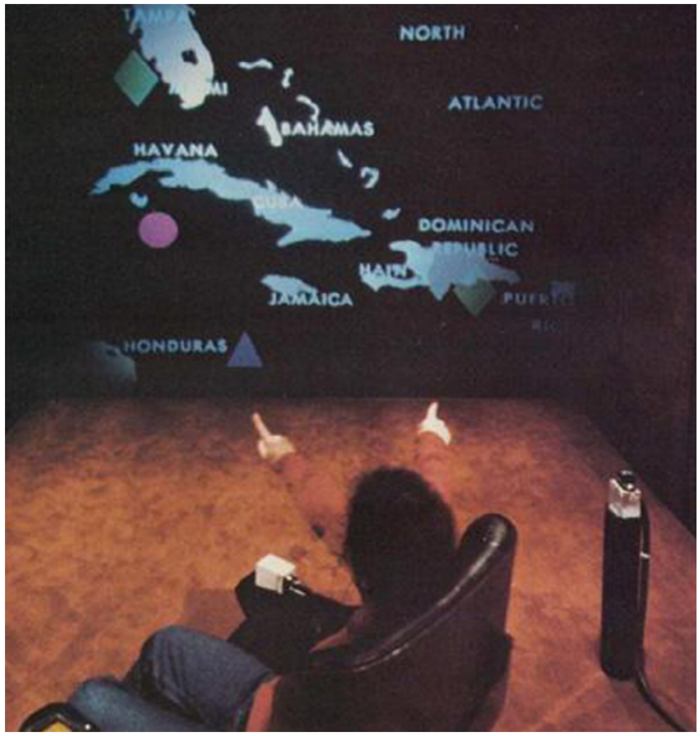

Figure 2-22: Bolts Put That There System (Bolt, 1980).

have more prominent expressive power, adaptability, and convenience [45]. With the aim to improve user interfaces of interactive robots with multimodality using speech and gestures, various experiments have been performed. The European Union funds one of the most prominent among them. The research was done by the Comm-Rob project (http://www.commrob.eu) and partially funded by the European Union. This research deals with the fusion of modalities and involves ordering a robot to perform tasks [46]. A Java-based FreeTTS API is used to convert speech into text. The robot used in the experiment is CommRob. The experimental results shown in the article depicts the robot as having very high accuracy. In one of the scenarios, it performed 100 percent accurately when the user called out to the robot robot while gesturing with both hands. They briefly mentioned that their fusion strategy uses a multimodal grammar that defines which terminal symbols (parts) of a particular command can be provided by which modality. In the following, the researchers present the fusion process with the example of the utterance "Go there [location]." The example chosen is a common phrase used in daily life for a complementary usage of speech (Go there) and gesture ([location] indicated by pointing) to produce meaning. They use grammar defined by the following production rules. The generated grammar then produces results by calling a robot function request (goThere(x,y)) [46].

$$phrase = verb - preposition - location$$

$$verb = "go" : \epsilon : request(goThere(e1))$$

$$preposition = "there" : Gpointing : \epsilon$$

$$location = \epsilon : G_(pointing) < x, y >: e1 =< x; y >$$

where,

verb, phrase, and location are non-terminals go, there, e1 (x,y) and epsilon are terminal

The command goThere(x,y) explained in above grammar is defined by verb, preposition, and location concatenated together. The resulting meaning is inferred as a command for instance (goThere(x,y)). The verb is used as "go" with the preposition "there" and location "Gpointing (x,y)" mentioning co-ordinates in the X-Y plane. The research performed using Comm-Rob uses a fusion model available in another research [47] published in Vienna, Austria. In this experiment [47], the authors discussed how to fuse various modalities, but they did not mention explicitly how they achieved fusion. The speech and gesture input are implemented using natural human conversation, and gestures are considered as good interaction tools. Kendon has defined gestures as voluntary and expressive actions of the human body used together with speech and perceived by the participant as a meaningful part of the speech [48]. A device extensively used for HMI for voice and gesture input is Microsoft Kinect.

The Kinect sensors comprise four different components, which are depth camera, color camera, microphone array, and tilting mechanism. In the research performed in 2013 in Switzerland, researchers fused gesture and voice using Microsoft Kinect. The architecture they utilized includes the Microsoft Kinect API, as shown in Figure 2-23, and captures the input from the user in the form of gesture and speech, which is fused later to draw meaningful operations from it. The results mentioned in the research show the multimodal selection performed better than its unimodal counterpart for total error versus user, namely, a number of errors performed during the experiment [49]. The multimodal system has better error handling capacity. The statistical parameter related to the mean average time of the multimodal selection calculated is 7.5 while the mean average time of the unimodal selection is 16.7, evaluated from Table 2.2. The results demonstrate that the Multimodal input is better than unimodal regarding the total error.

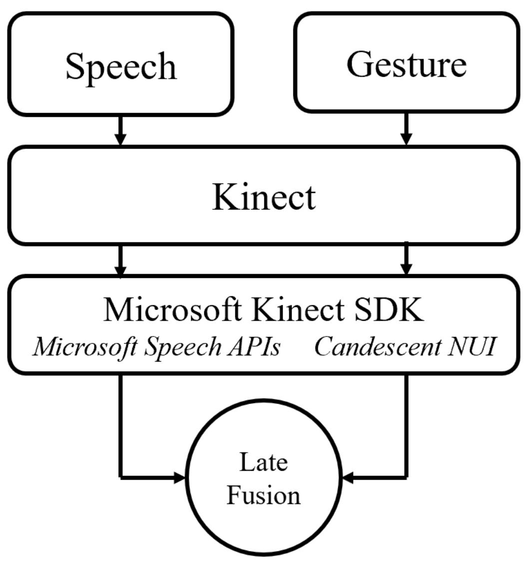

Figure 2-23: Implementing of Architecture for Integrating Speech and Gesture

Apart from Microsoft Kinect, to improve accuracy, cell phones can be used for automatic speech recognition and gestures for older adults by using mid-air gesture and voice commands to control the mobile device. For example, Apple's Siri, Microsoft's Cortana, and Google Assistant on Android devices can be fused with mid-air gestures. Though these voice assistants are quite accurate, some advancements should make them reachable to all segments of the society in terms of phonetics, especially Chinese, the Indian sub-continent, and Southeast Asian population, which pronounces various alphabets differently. The speech recognition database does not recognize elderly or shaky voice properly. Consequently, the performance of ASR (automatic speech recognition) reduces 9-13 percent when used by the elderly [13]. Some possible causes for such a deviation are alterations in the vocal chords, the vocal cavities, and the lungs, along with declining cognitive and perceptual abilities.

Table 2.2: Comparison of Unimodal and Multimodal Selection: Total Error per User

Type	1	2	3	4	5	6	7	8	9	10
Multimodal	10	3	13	4	5	8	4	7	3	18
Unimodal	2	3	22	19	4	51	7	5	20	31

ECOMODE is a new generation of low-power multimodal inetrface for cell phones, where the collaboration depends on vocal directions and mid-air gestures. ECOMODE's solution depends on two principle advancements namely mid-air gesture control set and a vision-assisted speech recognition framework. The solution to these issues, proposed by both these modalities are asynchronous in nature. The ECOMODE technology [13] combines the dynamics of the chin and the motion of the lips to achieve more robustness in the system. The state-of-the-art technology proposed will be designed to work reliably in uncontrolled conditions, particularly under excessive or low lighting and noisy environments.

Multimodal natural user interaction is performed for multiple applications to find the work progress concerning multi-application, multimodal interaction utilizing the Kinect gadget as a two-modular source. The research was carried out in the year

2012. The basic implementation deals with writing an application that understands gesture from Microsoft Kinect, along with consuming input through speech. The system architecture consists of three major components:

1. Receiving data from different modalities.

2. Compiling commands

3. Determining and selecting active apps that accept and perform commands [50]

The application is capable enough of fusing the data and deriving meaning from it. For example, if the system receives a gesture of Swing Right from the hand along with a vocal command, then it tries to move the Microsoft PowerPoint presentation to the next slide. Architecture is shown in Figure 2-24 . The advantage of this architecture is that it is not confined to a specific application. It could be used with any application running on a computer. A shortcoming of this research is that empirical data is not provided to support the claim, and there is no mention of accuracy or failure rate while using the specific application. The author concluded by referring to the future design of architecture in such a way that it will support the addition of new modalities, such as tension, pressure, facial expressions, and so forth.

The multimodal systems developed so far have a huge scope of improvement regarding grasping the speech from users of different ethnic groups. Ferreira et al., (2015) proposed a concept of socially-inspired rewards for improving the precision of a system. They are used to quantify reinforcement rewards, which are assigned to users according to how they have interacted with the robots based on vocal interaction. In this method, a potential-based reward-forming strategy instrument is joined with a sample proficient reinforcement learning algorithm to offer a principled structure to adapt to these conceivably chaotic conditions. The experimental setup comprises two live scenarios in which one is responsible for assisting a tourist in a given area called TownInfo, and the other is called the MaRDi dialogue system. MaRDi is responsible for performing a Pick-Place-Carry task in a human-robot interaction context, for instance, move the blue mug from the living room table to the kitchen table [51]. The MaRDi experiment involved a tightly coupled 3-D simulation software known as MORSE, while TownInfo worked as a virtual tour guide. Gesture and voice input were provided and fused to obtain semantics and match with a static knowledge base.

The recent article published on multilevel sensor fusion with deep learning deals with the fusion of information coming from different sensors. The design was implemented to achieve a trade-off between early and late fusion. At each level of abstraction, the different levels of deep networks are fed to a central neural network, which combines them into common embedding [52]. The fusion of audio and visual for evaluating the emotion of a human was performed by Shamim et al (2019), the results achieved are quite exemplary. The experiment receives accuracy of 96.8 percent, and it is fairly good in comparison with other experiments. CK+ database is used, the results achieved through FaceNet2ExpNet is good in comparison with deep sparse autoencoders and DNN using the same database [53].

The data fusion using various sensors and audio devices implemented in a hospital room to analyze the environment around the patient and His or her various needs. This experiment is used for academic purposes and provides detail insights for the College of Nursing students. The various devices used are localization sensors, microphone array, patient simulator, lapel microphones, and physiological wristbands. The experiment is quite a niche in its area, and its been implemented quite well for teamwork and its collaboration for a patient [54].

The application of data fusion techniques applies to all industries. The same idea is exploited to analyze the learning outcome of a session especially when students are supposed to take classes online. Their activity is observed throughout the class using click-stream data, eye-tracking, electroencephalography (EEG), video, and wristband data. The accuracy achieved was 94 percent using normalized root mean squared error implemented using prediction random forest technique [55].

The fusion of multimodal data used for rating prediction framework of consumer products by combining EEG signals and sentiment analysis of product review. The experiment uses Emotiv EPOC+ for EEG signal and reviews provided others customer in the form of text. The accuracy achieved is 71 percent. This experiment deals with the crude input provided by Emotiv EPOC and does not get into the minute details of the wavelet [56].

Chanoh et al. (2018) implemented a fusion technique called the dense map-centric SLAM method. It is based on combining multiple frames of a handheld LiDAR and compensate the remaining information to complete the image using Fusion. The Trajectory error in meters achieved using this technique is 0.076m from a length of 360 meter input, and the time frame required is 9.1 minutes. The experiment effectively reduces LiDAR noise by Surfel fusion [57]. In a bigger context, if the experiment is performed outside, global optimization may not be achieved.

In another experiment, the fusion of images captured by a camera and GPR/encoder data that are spatially evenly-spaced are captured and fused for subsurface transportation and infrastructure inspection. The proposed algorithms need to improve further for accuracy and speed, but in the ideal condition, the accuracy achieved is 98 percent. The experiment was performed on the bridge deck at the Ernest Langford architecture center at Texas A&M University to test the system [58].

Supervised learning is used as training data for a map application while capturing the images from the camera and annotated images. An experiment was performed to create inexpensive technology for marking road segmentations. It will help to create rules for maps used in autonomous vehicles. The accuracy achieved in the experiment is 75.04 percent. Moreover, the experiment does not include semantic classification of the road markings to retrieve the rules of the road [59].

For sensing human reliance on texture recognition, an experiment is performed using GelSight tactile sensor for capturing tactile images, which are further fused with vision using deep neural networks. In this experiment, the accuracy achieved is 90 percent by implementing a fusion method named Deep Maximum Covariance Analysis (DMCA). Using the algorithm based on DMCA, it is easy to calculate the perception performance of either vision or tactile sensing. The limitation of the experiment is temporal information is not included during the experiments [60].

The application of multimodal fusion was performed in surgery, while the surgeon is using MYO armband for gestures, and EPOC Emotiv for capturing the EEG signals and Microsoft Kinect for speech and capturing the body movements. The accuracy achieved by the experiment is 88 percent. Moreover, the device is too complex; it would not be easy to perform surgery while having two devices on the body. The experiment talks about excluding or minimizing the workforce required in the operation theatre. The idea is to read the surgeon's mind and provide the surgical tools required during the surgery. In a typical scenario, there are around ten professionals required in the operation theatre. If this experiment is implemented in a commercial environment, we will able to reduce the manpower required at the hospital [61].

The application of a robotic arm is used in calculating the depth estimation of metallic pieces. The robotic arm developed consists of Microsoft Kinect, lasers and mono-camera, and the system is designed to find featureless objects such as metallic plates, metallic connectors, and monochromatic objects. After the experiments performed, 95 percent achieved and 100 percent with the help of a human operator. The system is further improved by adding utilities and making it viable to be accepted by the industry [62].

Similarly, the application of multimodality is widely exploited in the health care industry. In an experiment, video and kinematics data used to perform surgical operations. The experiment claims to achieve the accuracy improvement of 15.2 percent using unsupervised trajectory segmentation based on a TCS-K approach. A SCAE network is used visual feature extraction from the input video [63].

The application of multimodal data fusion can be used in capturing cultural attributes, using a sensor simulator and a signal generator. The experiment claims to fuse attributes with heterogeneous information. It can learn new user attributes from distributed data streams such as human behaviors in different situations. The various attributes are talkativeness, extroversion, uncertainty, individualism, and personal distance. The limitation of this experiment is that, it is unable to find information on the user's current state (e.g., mood and satisfaction level, etc) [64].

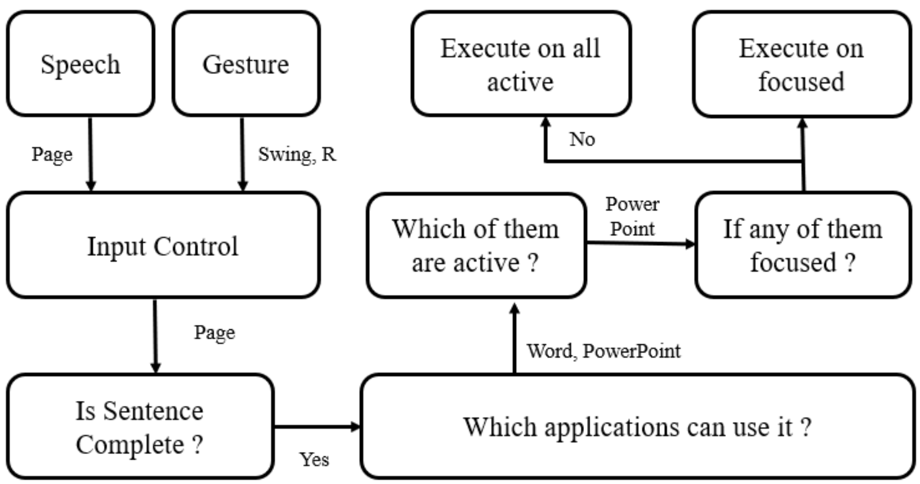

Figure 2-24: Two Applications Multimodal (Gesture-Voice) Example

2.2.2.1 Oculus Rift 3D

The next HCI device that we have explored is Oculus Rift 3-D goggles. The Oculus Rift goggles as shown in Figure 2-25 are a virtual reality headset developed and manufactured in 2016. The user pulls a helmet over his head, and suddenly, he is inside a virtual world that seems completely lifelike. The user can run around, fight, race, fly, and create ways that gamers (or anyone else, for that matter) have never done before. This is a great form to interact with your devices. It absorbs you into another world, and it makes users interact with the device in ways never thought of. The VR devices are becoming more and more popular as technology is advancing so much. Picture a set of ski goggles, but instead of miles of fresh powder, you are transported into space or underwater. The Rift accomplishes this using a pair of screens that display two images side by side, one for each eye. A set of lenses is placed on top of the panels, focusing and reshaping the picture for each eye and creating a stereoscopic 3-D image. The goggles have embedded sensors that monitor the wearer's head motions and adjust the image accordingly. The latest version of the Oculus Rift is bolstered by an external positional-tracking sensor, which helps track head movements more accurately. The result is the sensation that you are looking around a 3D world. Although it is nice to be able to leave reality for a moment, these goggles are very expensive. However, they are a good way to interact with a device in a virtual transportation to the unknown. This device is capable of having input from several modalities, including gesture and speech, simultaneously [65].

Figure 2-25: Oculus Rift 3D

2.2.3 Multimodal Systems without Fusion

Multimodal systems without fusion are defined as an application with multiples, but they are not complementing each in the completion of vague of partial inputs. A robot named Motherese is used to develop multimodal emotional intelligence. In this study, the authors were attempting to develop the concept of Multimodal Emotional Intelligence (MEI) [66]. As humans perceive the effect of voice, movement, music, and point light displays, the MEI robot accepts input in the form of voice and maps it to other modalities. The MEI robot uses various parameters to understand and express multimodal emotions that are defined by SIRE (Speed, Intensity, irRegularity, and Extent). The inputs used in implementing the MEI model were .wav audio files, a Microsoft Kinect for capturing gestures, and Flute for the music, as shown in Figure 2-26. The generation of emotional expression using MEI is implemented through the intensity and speed of speech. A voice capture with parameters in the ranges could be judged as displaying happiness or sadness accordingly.

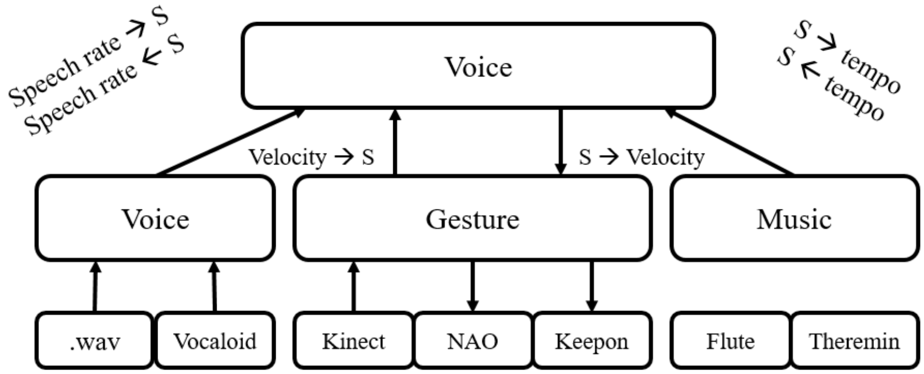

Figure 2-26: SIRE Paradigm for Experiments across Voice, Gesture and Music

The experimental results show the system is not robust enough to comprehend and analyze all the gestures properly; it sometimes gets confused between anger and happiness mentioned in Table 2.3; for instance, it has identified Happiness with 62 percent accuracy, whereas it identified sadness, anger, and fear by 0 percent, 19 percent, and 19 percent accuracy respectively. The researchers claimed that the system's accuracy is 63 percent in the first iteration and can be calculated by taking the mean of the diagonal values of Table 2.3. In the later stages, they reached accuracy of 72 percent [66] which we believe could be improved by training the system in a more diversified environment, such as, a twenty-dimensional confusion matrix.

Table 2.3: Confusion Matrix

Detected	Happiness	Sadness	Anger	Fear	p-value
Happiness	62	0	19	19	¡0.0001
Sadness	2	90	0	6	¡0.0001
Anger	55	0	43	2	¡0.0001
Fear	21	12	12	55	¡0.0001

Following a similar concept, a multimodal manipulator control interface was designed which uses speech and multi-mouch mesture recognition. The research deals with managing a robotic arm with touch and gestures [67]. The degree of freedom for a robotic arm is seven, which is controlled using rotate, open and close commands for the gripper. Per the claim, the robotic arm could be used by novice users, and they could operate the robotic arm easily. The interface recognizes five types of touch gestures: slide, open, close, clockwise, and counter-clockwise. In control mode, open and close are used specifically for the gripper. While the left and right gestures are used to move gripper left and right respectively, the prototype has been developed using the seven degrees-of-freedom robotic arm, using a manipulator, which includes a laptop with a touchscreen [67]. The robotic arm used to have six joints and a gripper. Figure 2-27 illustrates the system, comprises of a laptop with a touchScreen, a CAN-to-USB adaptor, a USB headset, and the robot with six joints and a gripper. The touch gestures mentioned in Figure 2-27 explain how fingers should be used while performing slide, open, close, clockwise and counterclockwise operations to control the iARM. T. Oka et al. (2005) have [67] implemented the multi-touch gesture recognizer, which detects and understands the gestures on the touchscreen. The system was designed for recognizing a aprticular grammar, which includes three rules and thirty-five words. The interface recognizes multimodal commands using spoken language and only one contact point over the panel. When the system receives a multimodal command, the system sends velocity and position commands to the manipulator, moving it to the defined location using three arms. The pilot study reveals that new users can control the manipulator [67]. They can easily pick up, rotate, and replace objects using gesture and multimodal commands. The limitation of the results shows that the user would not be able to operate effectively in a rotational mode.

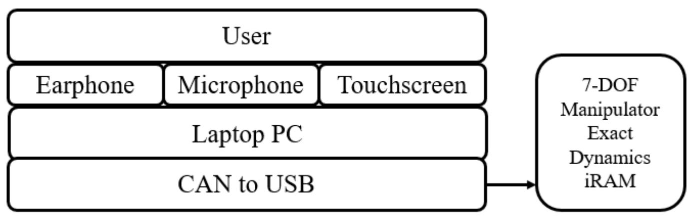

Figure 2-27: Hardware Configuration

Krum et al. (2002) implemented a multimodal navigation interface [45] using speech and gesture for a whole 3-D visualization environment. Virtual Geographic Information System (VGIS) is used as a multimodal interface, which provides a set of 3-D navigation of the globe. Speech and gesture multimodal control is implemented in Earth 3D Visualization Environment; research was done in 2002, which aims at reducing the error rate in multimodality in comparison to unimodal component interfaces. In noisy environments, users can rely on pen and gesture input, while the differently-abled users can use speech. Those users who do not have a clear accent or shaky voice will prefer to use gesture and pen. One of the important aspects of multimodality is that the user may not have peripheral devices, like a mouse or keyboard, to provide the input, and he or she may mostly be occupied using his or her hands for moving around the display most of the time. It is important to understand the limitations of multimodal speech and gesture interfaces rather than comparing performance with other interfaces. The VGIS systems allow the user to travel from the orbital perspective of the entire globe, which displays 3-D building models and sub-meter resolution images of the earths surface. Navigating an extended 3-D space in VGIS brings more challenges to the applications:

1. Including scale, seven degrees of freedom must be managed.

2. In a virtual environment, good stereo imagery must be maintained.

3. Navigation methods must work at all spatial scales.

Krum et al. (2002) were able to address 1 and 3. The implementation part of multimodal interfaces used a variety of hardware in a desktop PC, a laptop, and a Fakespace Virtual Workbench powered by an SGI Onyx2. Voice recognition was performed by IBM ViaVoice in which speech utterances are converted into text and sent as commands over the network; sample commands are shown in Table 2.4. A gesture pendant is worn on the human chest which has an LED, and it captures the hand movement. The camera has an infrared filter which is having the best feature; it avoids other light sources from interfering the image. The speech commands, mentioned in Table 2.4, are then translated to multimodal interface commands, based on the mapping, which enables the model to render the requested image, building, etc. Navigation commands, as shown in Table 2.5, are available to navigate in x and y directions. The system works in three modes.

These modes are the orbital mode, walk mode, and fly mode. In walk mode users can constrained to the ground, orbital mode presents a third-person point of view, which always looks from down to above, and the fly mode presents helicopter-like flight. The performance of the system is evaluated on the metrics mentioned below.

1. Gesture recognizability and responsiveness: how accurately and quickly the system recognizes gestures and responds

Table 2.4: Sample of Recognized Speech Commands

Type	Commands
Modes of Navigation	Orbit, Fly, walk
Continuous Movement	Move In, Out, Forwards, Backwards height
	Move Left, Right, Up, Down height
	Move Higher, Lower
Discrete Movements	Jump Forwards, Backwards height
	Jump Left, Right, Up, Down height
	Jump Higher, Lower
Direction	Turn Left, Right height
	Pitch Up, Down
Speed	Slower, Faster, Stop

2. Speed: efficient task completion

3. Accuracy: proximity to the desired target

4. Ease of learning: the ability of a novice user to use the technique

5. Ease of use: the cognitive load of the technique from the users point of view

6. User comfort: physical discomfort, simulator sickness

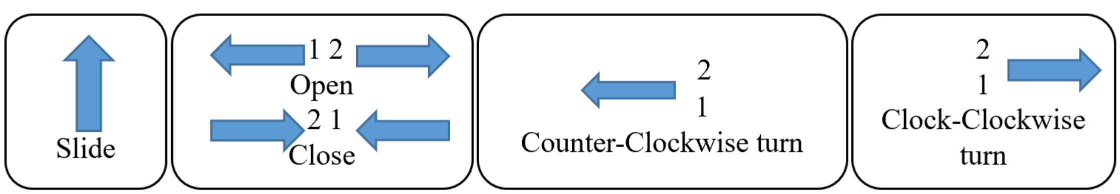

Figure 2-28: Touch Gestures

Table 2.6 summarizes both commands. A novice user must go through the training of fifteen to twenty minutes to understand the system and learn commands. Recognizability and responsiveness of voice recognition lags were factors in the performance of the users. Also, users would sometimes have to repeat commands, but overall the

Table 2.5: Recognized Gestures

Vertical Finger Moving Left: Pan Left
Vertical Finger Moving Right: Pan Right
Left Finger Moving Up: Zoom Out
Left Finger Moving Up: Zoom In
Right Finger Moving Up: Pan Up
Right Finger Moving Down: Pan Down
Open Palm: Stop

system is capable enough to understand most of the times. With regard to speed and accuracy it took a total of 10.1 minutes to complete one task, which is too long, and the user had to utter 50 to 100 spoken commands, whereas the mouse interface just took 3.5 minutes [45]. The accuracy of the system is pretty good with a very few chances of error, and that too while adjusting more detailed movements. If we talk about ease of learning, ease of use, and comfort, the system is capable of addressing all these aspects. Some users prefer keystrokes rather speech and gesture. There are certain commands which are not present in the system, or if the user provides a wrong command, the system will prompt "Command not found." The designed system has an ease of use which allows the users to move higher command with an increased rate of motion but a decreased rate of motion in case of lower motion command, which could be confusing at times because we are addressing both speeds, up and down, with a single gesture, just by altering the speech input. The system is not that comfortable as it fatigues the hands while using the system.

Speech and arm motion were explored in a multimodal context by Bozkurt et al. (2016) which was discovered in Istanbul, Turkey, in 2016. The goal of the research was to implement machine learning in multimodality. In virtual environment designs, gesticulation is an important concept introduced in the paper. Gesticulation deals with adopting and emphasizing the human-centered aspect, which is missing in virtual environments. The study explored a programmed combination of motion in synchrony with speech and joined with nonverbal correspondence segments into virtual character segments. The study deals with the feature extraction of unimodal clustering using both semi-supervised and unsupervised forms of clustering. For semi-supervised learning, a pool of gesture input is provided using the Hidden Markov Model, while in unsupervised learning, a large-scale multimodal dataset is used [68].

For the unsupervised learning experiment, the researchers made a twenty-minute recording of five different native users (data are shown in Table 2.7), all of them being Turkish in origin. The speakers wore a black suit with fifteen color markers and a microphone, placed close to their mouths, and synchronized with their speech. During the experiment, the users were not instructed on how to provide the gesture input specifically. The number of gestures collected by all the users is shown in Table 2.7. These gestures are analyzed by using the semi-supervised learning technique, which shows that each user performed seven unique gestures.

Further research regarding gesture and voice interaction with interfaces was carried out at the University of Glasgow in 2012 in the research by Rico et al [69] . The paper dealt with the issues of social acceptability and user perception as they related to multimodal mobile interaction techniques. The evaluation of social acceptability explored performance regarding audible and visible interactions, including how the user perceived the interaction and how comfortable they were while using the device. The exact scope and definition of user experience are still debatable, but while designing handheld devices, the designer should take into account the individual thoughts of users and their feelings and reaction to an interface. To understand the users behavior, an experiment was carried out with sixteen gestures and sixteen voice commands. Voice and gesture were chosen because of their highly visible and audible nature. The modalities were examined on an individual basis, rather than grouped together. The gesture was classified into four subcategories, namely, emblematic, device based, arbitrary, and body based. The most widely accepted gestures are deemed emblematic, while device-based are those who are involved in directly manipulating a device. Arbitrary gestures are defined as those set of gestures which were defined previously as emblematic and device based. Body-based gestures are those in which direct physical contact with the device is not involved. Body-based gestures work with external sensors and capture the bodys movement. The voice commands used classified into three categories: speech, command, and non-speech. The command is a type of input in which user says short words, for example, call or lock and so on, while speech input is short commonly used phrases and non-speech input including noises which are used in everyday life (the gesture details are mentioned in Table 2.6). The experiment was carried out such that half of the users could use gestures while the rest were told to use voice commands. After the experiment, an interview was conducted in which a worksheet was provided to collect feedback on what the users felt regarding input preference, locations where these inputs might be used, tasks where these inputs could be used, and so on. A total of nineteen participants were involved in this study, with the majority ranging in age from eighteen to twenty-nine, while two local community members were between the ages of seventy and ninety-five. The results show that device-based gestures were preferred over other gestures, body-based gestures ranked second, arbitrary gestures were least acceptable or preferred, and emblematic gestures were second least acceptable. In speech input, commands are most widely used and had high acceptability in comparison with speech and non-speech commands. Other research, this time focusing not on user interaction with multimodal systems but on how multimodal systems interact with desktop applications, was performed in 2013 [70].

Continuation of a part of previous work, reported by Niloas et al. (2012), the research behind the 2013 paper, was performed by a joint group of the Applied Informatics and Multimedia Department, Greece, the Electronic and Computer Engineering Department of Brunel University, UK & Medialogy Section, Copenhagen,

Table 2.6: Gestures and Voice Commands, by Category

Gesture	**Category**	**Voice**	**Category**
OK Gesture	Emblematic	Say "Close"	Command
Money Gesture	Emblematic	Say "Open"	Command
Peace Sign	Emblematic	Say "Call"	Command
Shrugging	Emblematic	Say "Lock"	Command
Device Stroke	Device-Based	Say "I'm Fine"	Voice
Device Shaking	Device-Based	Say "Bad Weather"	Voice
Device Flick	Device-Based	Say "That's Nice"	Voice
Device Rotation	Device-Based	Say "So Busy"	Voice
Upright Fist	Arbitrary	Humming	Non-speech
Hook Finger	Arbitrary	Buzzing	Non-speech
Sideways Fist	Arbitrary	Say "Chh"	Non-speech
Open Palm	Arbitrary	Doo Doo Doo	Non-speech
Shoulder Rotation	Body-Based	Say "Psst"	Non-speech
Wrist Rotation	Body-Based	Whistling	Non-speech
Foot Tapping	Body-Based	Clicking	Non-speech

Denmark. As in the previous paper, the researchers used Microsoft Kinect with multiple sensors to scan the face completely and precisely. In this paper, a multimodal natural user interface system, based on real-time audio, video and depth processing, was demonstrated. To illustrate the concept, they have four possible scenarios:

1. Login via face detection system, which we have seen recently in Windows 10;

2. Application selection via object detection-recognition;

3. Authorization control according to log in and data, and;

4. Application operations.

The input devices consist of an RGB camera, and depth and audio sensors, with each device working independently without hampering the rest of the system. The system architecture implements a multimodal system based on natural user interaction as shown in Figure 2-29. The multimodal process involves face detection, objection detection, speech recognition, and gesture recognition. The face detection is completed in two steps. first, the joints of the head are tracked by the Microsoft input received. This cropped image does not provide sufficient data for user recognition since joints are not always stable which leads to the use of a two-way authentication process. A face detection algorithm is used to extract recognition data from the cropped image as shown in the Desktop Login Flow Diagram in 2-30. This combined method is then used in the future, that is, authenticating the user when they attempt to login on subsequent occasions [50].

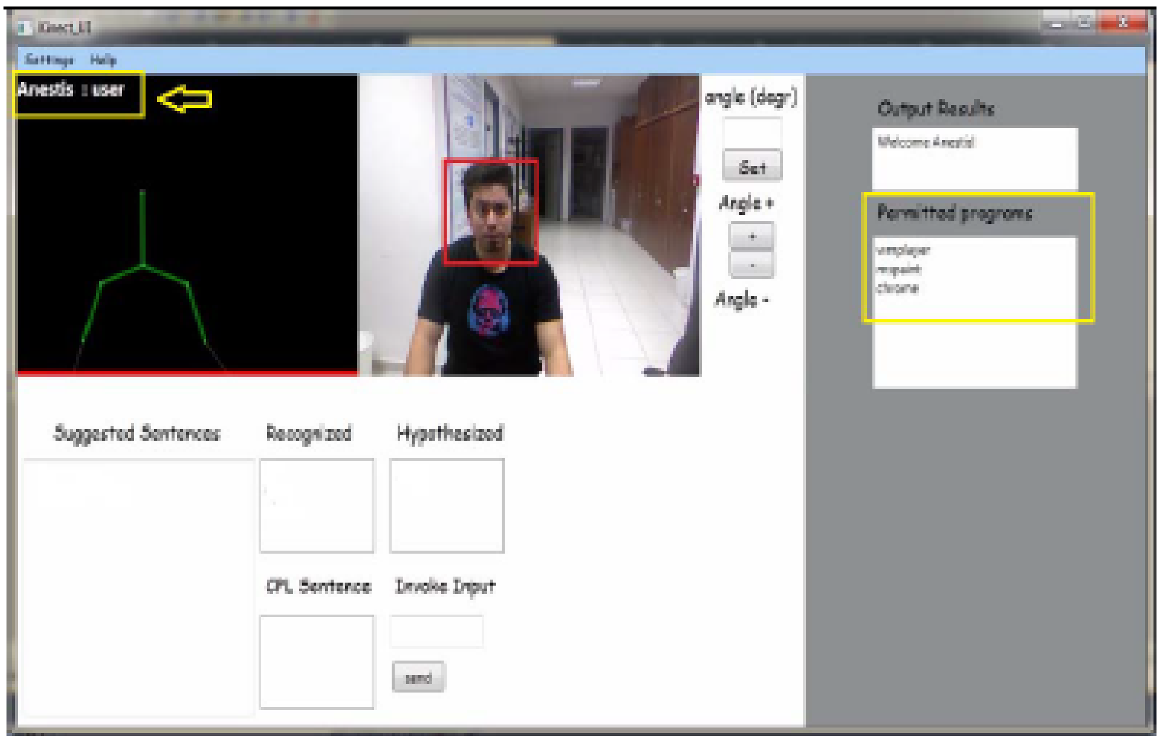

Figure 2-29: Simple User Rights Login

There was no empirical data provided, but in all four phases, experiments were performed, and the user was able to log in, use an application, and perform operations using Google Chrome and MS Paint [44]. Another niche concept is explored, where intention recognition is used, a fairly novel idea, which includes intention recognition in conjunction with multimodal systems in the year 2015 via research at the Electronics and Telecommunications Research Institute in Daejeon, South Korea, resulted in a

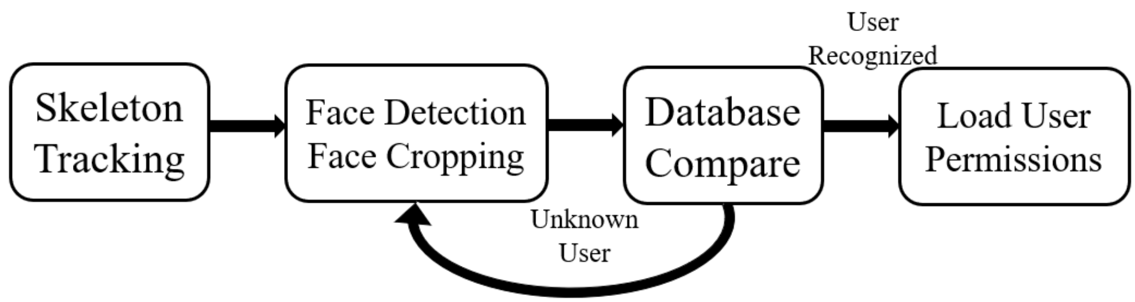

Figure 2-30: Desktop Login Flow Diagram

paper entitled "Multimodal Data Fusion and Intention Recognition for Horse Riding Simulators." The research developed a system that gives a user the feeling of riding a horse and attempts to teach the user the skill of horseback riding. As it is not possible or feasible for everyone to learn horseback riding on an actual horse, this prototype enables users to learn and ride within a simulated environment. The proposed system consists of multiple data acquisition components, a feature extraction component, and a data fusion component. The system can increase realism for the user by enabling riders to perform interactions similar to ones they would perform while riding a horse.

The hardware consists of a multimodal user interface, multiple cameras, microphones, and other sensors to capture the users natural speech and movements. Three kinds of sensors were used namely contact, auditory, and two visual sensors. One contact sensor was mounted on the body of the simulator which senses that the user is riding the horse, while the user wore additional contact sensors. The auditory sensor was mounted on the helmet of the horse-riding simulator and captured the voice commands from the user. Two depth-sensing devices were used for capturing visual information. The contact information captured by the contact sensors included balanced sitting, drawing or pulling reins, spur, whip, and so forth. Using "Gesture and Speech Control for Commanding a Robot Assistant," the researchers performed experiments used a robot called ALBERT [71]. They conducted experiments using gestures by considering the heuristics of hands by utilizing a webcam available on the robot. The experimental results show 95 percent correct recognition of hand gestures displaying yes and no (thumbs up is considered a yes while thumbs down is a no). Verbal Input ViaVoice, which is a speech recognition software offered by IBM, is used. This research does not deal with fusion at all; the experiments performed separately.

Table 2.7: Gesture Phrase Distributions Per Recording

Red. Id	g1	g2	g3	g4	g5	g6	g7	Total Count	Dur (s)
1	52	64	9	22	1	0	19	167	239
2	20	40	1	8	0	17	6	92	167
3	22	49	1	23	10	21	40	166	265
4	53	60	15	20	4	18	20	190	347
5	2	45	1	0	0	0	0	48	155
Total	149	258	27	73	15	56	85	663	1173

In another experiment, intention recognition is the niche concept explored while using horse-riding simulators. This idea is completely new and has not been discussed elsewhere in the literature surveyed herein. Intention recognition is implemented by defining a class that stores every input received and updates the database whenever it receives a new input. For every action, an intention class is defined. For example, the balancing intention class captures the actual position of the user, and the exercise maintenance intention class corresponds to the leg release or bridle release actions. The main intention class includes strength information expressed through the action as show in Figure 2-31. Once all the data is collected, it is compared with data from the intention database. Based on the results, the system recognizes the intention of the user. The researchers have included multimodal data fusion but have not shared any empirical data about the accuracy of the system [72].

Building a multimodal human-robot interface, a paper published in 2001, talks about how to build a system which can accept multimodal inputs. The author talks about personal digital assistants as a form of input apart from speech and gesture [73]. The media center application designed, while considering the requirement of the differently abled users, and it exploits minimal hardware, namely. PC (Athlon X2 3800) running the media focus server programming, a cell phone (Nokia N95) for interfacing with the client and running the customer programming, a remote get to indicate associate these together, and a superior quality forty-inch advanced TV showing the UI.

(a) RGB (b) Depth (c) Mask (d) Skeleton

Figure 2-31: Sample Cues of the Multimodal Gesture Challenge 2013 Data Set

Skubic et al. (2004) designed a system called Spatial Language for Human-Robot Dialogs which enables the robot to analyze surroundings and managing things around it. In an example, it has been illustrated that the robot can analyze things around it and respond verbally where things are located. Spatial language has been used to define the geographical locations of objects lying in a room. The idea has been discussed though it has not been implemented and no statistical data is provided in the paper. An example is illustrated in the dialogue below [74].

1. Human: How many objects do you see?

2. Robot: I am sensing seven objects.

3. Human: Where is the box?

4. Robot: The box is behind-right of me. The object is close.

Multimodal Media Center Interface based on speech, gestures, and haptic feedback was designed in the year 2009 in Finland. The proposed solution consists of a multimodal media center interface based on speech input and haptic feedback. The system architecture contains a zoom-enabled, context-focused GUI, tightly coupled with speech input. The core idea of this research is to make a substitute system for trivial digital home appliances, such as, remote controls, game controllers, and so on. The author has argued that many of these interaction devices are not user-friendly. A classification is made based on the visual ability of the user. For a blind user, speech and haptic inputs are sufficient for accessing the information, while a zoom-enabled GUI is proposed for visually impaired (low-vision) users [75]. The application was developed using C and ran under Windows XP. The application consists of speech recognition while the mobile device can interpret the gesture, speech, and haptic feedback through a GUI. The mobile technology was based on a native Symbian application while the GUI and main logic used MIDP 2.0. The media center provides an electronic program guide which enables the user to access complete digital television content. The system consists of two graphical displays: a television and a mobile phone display. The first GUI on the television uses a matrix format to display the interface, explain its usage, and presently available content. One other proposed solution is to mount a wireless microphone instead of using a mobile phone for physically challenged users. It is unknown when or if the proposed system will be implemented. Assistive Robots for Blind Travelers is an ongoing project at the CMU Robotics Lab in which a robot is attempting to guide visually impaired people through an urban environment. For individuals with disabilities, transportation remains a major barrier for living a quality of life. With the advent of robots, it could be argued that their life would be much easier, especially with the usage of smart buses and shuttles. The differently abled people can live a healthy life, but they cannot drive, eventually makes their life tough. The visually-impaired can use the systems based on physical, verbal and digital input defining the foundation of human-robot interaction. The objective of this project is to enhance the safety and independence of visually impaired travelers.

The implementation involves the following three pieces [76]:

1. Rathu Baxter: Rathu Baxter was originally designed to assist human manufacturing settings
2. Mobile Robots: Mobile Robots can enhance the navigation experience of blind and visually impaired travelers in urban environments.
3. NavPal: NavPal is a smartphone app to give navigational assistance to blind adults as they move around unfamiliar indoor and outdoor environments. As it is early in the implementation phase, not many details are available.

Human-speech perception is a multimodal process which provides higher knowledge resources such as grammar, semantics, and pragmatics. The information source, which is used in the presence of noise, is lip-reading or also known as speech-reading. Automatic speech recognition (ASR) is a very active research area for several decades, but despite the fact many teams are working on it still would not able to compete with the performance achieved by human ears: the results achieved by ASR systems are far lacking from expected results. Most state-of-the-art ASR systems use acoustic signal only and ignore visual speech cues. Therefore, they are susceptible to acoustic noise, and all real-world applications are prone to error because of some noise in the background [77].

Implementing the concept of multimodal human-robot interaction framework, a personal robot was designed at Universidad Carlos III de Madrid. The architecture, used in developing the prototype, is Automatic Deliberative, which incorporates an emotional control system (ECS). The Automatic Deliberative (AD) architecture is based on a human psychological model. In the deliberative piece, the robot can do tasks such as planning and word model management, while the automatic piece pertains to reactive and sensory skills. An emotional control system is added to this Automatic Deliberated Architecture, as shown in Figure 2-32. By using this architecture, the researchers would be able to train the robot in skills such as greeting, face recognition, user identification, dialogue, audiovisual interaction, non-verbal visual expression, and dancing. The system designed has been named Maggie. There are three modalities that the researchers have proposed for use: visual, voice, and audio-visual mode.

Visual mode is enabled via the Proxemics and Kinesics expression control; the voice mode uses the Text to Speech library VTxtAuto [78] (VoiceText 1.0 Type Library) to generate speech from text. For the audio-visual mode, images and sound expression input is provided through a tablet PC, and Maggie understands it by utilizing Pure Data and Graphic Environment for Multimedia (PD-GEM), which is an open source audiovisual software tool. In this research article, the authors have not presented any empirical data that quantifies the accuracy of their robot. As per our understanding, it is just a hypothesis that explores the possibility of using multimodality in robotics. The proposed framework could perhaps be used to implement game scenarios and choreography programming.

Multimodal input from a robotic arm is implemented to perform grasping, unscrewing, and insertion tasks on a Barrett's robotic arm. The inputs involved are multimodal sensory signals, and it achieves 80-90 percent while performing the tasks. The drawback of this system is that it only worked on those trajectories that have

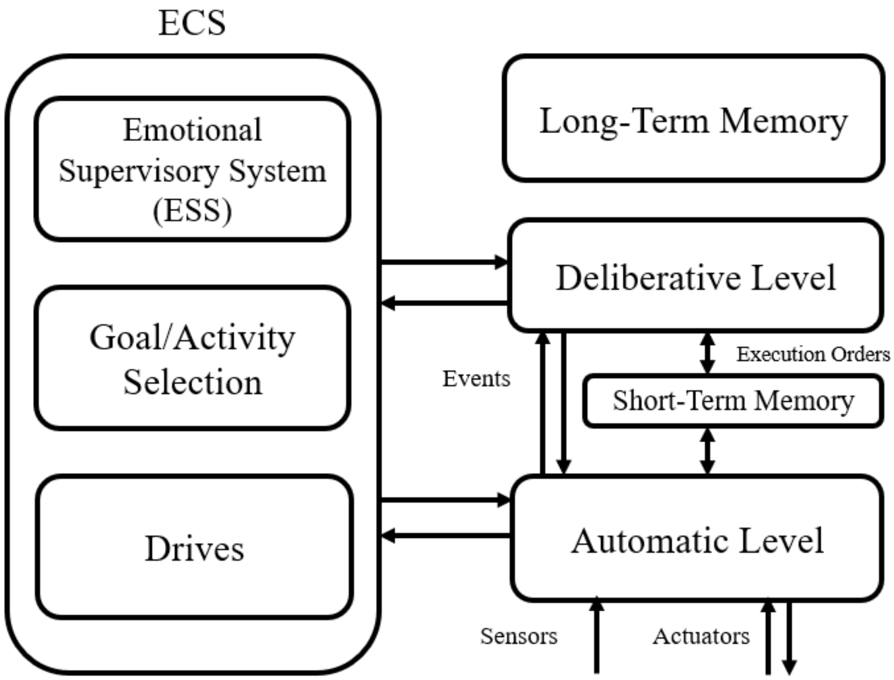

Figure 2-32: Automatic-Deliberated Architecture

been shown to the robot earlier, which implies it works in a constrained environment [79].

So far, we have seen various applications of multimodal data fusion, including automation in health care domain. Another interesting experiment was performed in Germany in the year 2018. A device is used with sensors to capture multimodal data for the detection of users motion intention and its assimilation into the exoskeleton control system while climbing the stairs or walking. The systems claim to achieve the average accuracy of 92.8 percent using the Hidden Markov Models (HMMs). While analyzing the system its been observed the placement of IMUs, and force sensors are used for capturing the data. Moreover, the details of fusion are not explained explicitly. The algorithm used is unable to predict a deeper analysis of the latencies for different motion transitions [80].

The application of multimodal systems varies from assigning tasks in an industrial environment to help humans in climbing the stairs. Along similar lines, a concept was introduced called mixed reality, in which Microsoft HoloLens is used to perform Pick-and-Place tasks on things placed on the ground. The experiment has achieved an accuracy of 93.92 percent, but it will only perform simple trajectories. The experiment demonstrated Microsoft HoloLens camera is mapped with the motors of the robotic arm and perform actions. The idea is not very new, but the implementation needs to achieve the accuracy for complex trajectories [81].

Table 2.8: Human Sensory Modalities Relevant to Multimodal Fusion

Modality	**Example**
Visual	Face Location
	Gaze
	Facial Expression
	Lip Reading
	Face-based identity (such as age, sex, race etc.)
	Gesture (head/face, hands, body)
	Sign Language
Auditory	Speech-input
	Non-Speech audio
Touch	Pressure
	Location and selection
	Gesture

2.2.3.1 Computer

One device that has continued to be upgraded and improved through the years since it was first created has been the computer. One can argue when it was first created which has led to the discovery of two possible starting points back in 1622 when William Oughtred created the first blueprints for a very crude-looking computer. One the other hand, in 1833 and 1871 when Charles Baggage created the first computer that resembled our modern ones, and it was called the Analytical Engine. With the computers going as far as the 1600s, it is insane to see how advanced and powerful it has become when back then one could only do the simplest of functions or commands with it. In current times, these have become one our staple points as so many people use them these days, not just for the latest tech but also for the internet and what we can do with it. Also, how computers were mainly used for research or creating new software and not for the personal joy of others till many years later now where almost an average American family will have some form of a computer as shown in Figure 2-33 in the household. Then there is the issue of how teachers are using them to help teach younger students, as since the internet has more to learn whereas it also has flashy things that can get young ones attention while learning new things. As to how far we have come with technology, it also comes with new threats in the digital zone as people can hack into bank account, and steal a whole persons life away with a single click. The fear of being hacked virus making its way into a computer or device leads their creators to try to develop the latest software defenses to protect them. While some can hold hackers and viruses at bay for some time, the other side is also trying to improve their methods to ruin people's devices or lives knowing how addicted modern people are to them [82]. PCs accepts input from various peripheral devices such as a mouse, keyboard, microphone, touchpad, webcam, and fingerprint scanner. All of them work independent of each other, which leads to conclude that a personal computer is a multimodal device without fusion.

2.2.3.2 Cellphone

A very common device in our modern times that has upgraded since the very simple phones from the 1900s is the cell phone. The very first cell phone was created for the world in 1983 when the company Motorola launched the DynaTac 800x for only $4,000 one could get a portable phone with a battery lifespan of 30 minutes, but back then this was revolutionary as it enabled people to walk and talk without being limited by a wire or cable. In modern times, the common cell phone has gone

Figure 2-33: Personal Computer

through so many phases by different companies that it is impossible to list them all, but nowadays they are mainly known by the best shape to hold in your hand or pocket, the best camera, the fastest data, and the newest and most trendy phone in that year. Some people would say that cell phones are in pretty much everyone's hands, and the companies do not show any sign of slowing down in their race to create the best phones and make the most money. The older phone are less fragile, whereas new phones are prone to corruption or are very fragile. That is one of the main reasons many Android users do not want to get an iPhone they can break very easily, and many are just slightly upgraded copies of their past versions. Many iPhone users love the smooth feeling and fancy covers they can buy for their phones, but the majority do end of chipping or shattering the screen, which are expensive to fix [83]. Sadly, many people are easily drawn into the buying circle of these phones as commercials usually depict some person with the companys latest phone having a good life or being the center of attention while taking pictures. Too many people, especially teens, are easily drawn into buying them to follow the trend, and this will continue until they either run out of money or the company stops producing phones. Some companies, though, will ask people to take surveys on what they want from their phones and try to meet demand so they can profit off them. Surveys indicate that what people want most from phones: a better look, faster access to internet, a nicer feel, and more storage. So companies will try to meet demands, and some do a pretty good job of upgrading their devices: others will keep the same format and make one minor upgrade and then sell the phones for a higher price. The latest cell phones in 2019 are capable of accepting a user's input from touch and speech. The idea of implementing error-prone device is still far away; on the contrary, most cell phone is keep track of human activity throughout the day, which is the biggest danger to the freedom of the user. The google maps application tracks all the location visits, and even if the user disables them, Google will keep the data for a month before deleting it. The health app captures the number of steps walked in a day and how many stairs a person climbs. Moreover, it makes it more susceptible to cyber attacks, not only in terms of money but also loss of information. Companies advertise their products based on lifestyle, for example, a user may have an issue with blood pressure; the person's cell phone will capture his or her details, and pharmaceutical companies will start offering insurance and medicines.

2.2.3.3 Apple Smart Watch

The advent of smart watches has brought revolution into our lives; millennials and teenagers are comfortable using the latest technologies. The Apple smart watch accepts touch-based input from the user. The new Apple smart watch 2-34 has many different features. These watches have features such as GPS, a heart sensor, and a speaker. The smart watch can access applications on your iPhone, such as messages and the camera. These watches are linked to your iPhone, which makes it convenient to answer phone calls and messages. Apple watches start at $399. Apple watches can be put on many different kinds of bands made out of leather, metal, or nylon. The new Series 4 watches are a little bit larger and a little thinner than the previous models. The new Apple watch has up to eighteen hours of battery life and can be wirelessly charged. These watches are great for notifications and phone calls. The new version also has Bluetooth built in and can pair with Bluetooth headphones and play music straight from the watch. Emails, phone calls, and text messages are easy to respond using this watch. Voice commands are also available, which make sending messages easy. Alarms can also be set, and the watch rings and vibrates just like a phone. This watch does it all, but it is a bit pricey for a watch. Siri responds to most common questions on the watch as well but does take some time to receive info from the phone. The Series 4 watches can come with either GPS or GPS and cellular data [84].

Figure 2-34: Apple Smart Watch

2.2.3.4 Microsoft Modern Keyboard with Fingerprint ID

Since we first began using PCs, one of the main ways we could interact with computer and input information through a QWERTY keyboard. We still use QWERTY keyboards just as frequently, although there have been some slight changes to them over time. There are ergonomic keyboards for comfort and prevention of carpal tunnel for those whose spend a lot of their day typing. Some keyboards connect wirelessly to your PC, eliminating some of the wire clutter. Backlit keyboards make it easier to see, especially if you are typing in the evening or where there is low light. One can get a keyboard that has a trackpad on it, eliminating the need for a mouse. Many of the keyboards, especially those made by Apple, are very thin. Many keyboards now have customizable shortcut keys. This Microsoft keyboard as shown i Figure 2-35 stood out more than the others because it has Bluetooth and a USB connection for recharging the battery. It was designed for comfort; it also has an added feature that is not often seen. It has biometric security included with a hidden fingerprint scanner for an extra secure but easy login option if you are running a Windows 10 operating system on your PC [85].

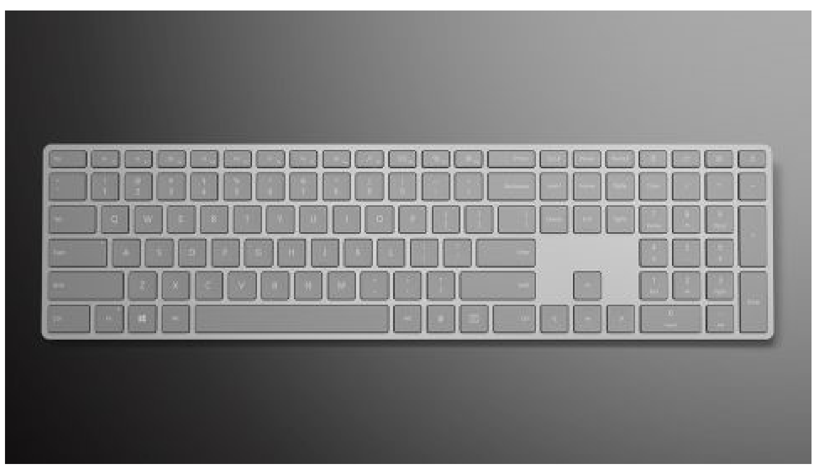

Figure 2-35: Microsoft Modern Keyboard

2.2.3.5 Autonomous Vehicles

Most of the smart vehicles of this era have capabilities such as communication with other vehicles, communication with the infrastructure (traffic signals and traffic update), GPS, sensor-driven decision making, etc. Connected autonomous vehicles (also known as smart cars) are driverless, capable of making their own decisions based on data from various sources (with little or no input from the user) and avoid obstacles that come in its way without causing any discomfort to the passengers or the other cars near it. These cars gather data from a myriad of sensors, the internet, roadside infrastructure, GPS, and so forth, and it is fed to the driving model which makes crucial driving decisions. Communication plays a significant role here, as most of the data coming to the smart car are from other smart entities, such as other smart cars and roadside infrastructure.

The self-driving car has both intrigued and terrified people over the years because of technical glitches observed and accidents that have happened on the road. If we move to the history of autonomous vehicles, one the very first autonomous cars was invented in 1925 when inventor Francis Houdina created a radio-controlled car that could start, shift gears, and perform most of the functions available that time while Houdina never touched the wheel. From 1925 to 2019, the progress of self-driving cars has continued to improve, and many hope they will become a reality without fear of the AI taking over and running over people. Sadly, in 2014, the first self-driving car fatality occurred when a Tesla tester died when the car hit an eighteen-wheeler. This has sparked a huge debate on whether people should continue to fund these projects or just let them fade away. However, many argue that if given enough time, the dream of autonomous cars can become a reality and may prove far safer than regular cars. One argument was how it could help people who were intoxicated and unable to drive can simply get in, and the vehicle will take them home without fear of crashing. Many creators and developers of self-driving cars want to keep pushing people to trust them and help them bring these futuristic vehicles to life. Surprisingly enough at CES 2018, where they announced the latest cars to be released or teased for the upcoming years. One company by the of name Nvidia revealed an autonomous car named Xavier where they shall incorporate AI into it. Many people who attended the event were very excited to see the self-driving car quest still going and now looks very promising with our current technology, many hope for the satisfaction of relaxing while the car drives them wherever they need to go. The company has also put out a teaser on how users can program the car either by typing in their destination or speaking the location to the google maps or any other navigation system. [86]. The first autonomous vehicle we are going to be discuss is Tesla.

Tesla

Tesla is a self-driving car made by Tesla and Elon Musk. Tesla was originally founded in 2003 by a group of engineers. Tesla was made not only for revenue purposes but also to motivate and influence the consumers to use zero-emission cars, which is a topic of great interest these days. With global warming and rising gas prices, more people want to transition to electric cars, but Tesla as shown in Figure 2-36 is not only an electric car but also self-driving. With this type of technology, HCI is inevitable. Just because the car could drive by itself does not mean a human is not important during the interaction process. A human is needed to tell the car where needs the human to pay attention at all times to avoid accidents from happening. Self-driving cars such as Tesla have multiple sensors place all around the car that help the car understand the environment so it can steer itself appropriately. The car has a high-precision, digitally controlled electric braking system, twelve long-range ultrasonic sensors, a forward-looking camera, and forward radar. The ultrasonic sensors are placed around the car so they can sense sixteen feet around the car. They sense when something is too close and also used for lane changes. So far, the interaction is using the touch screen-based tablet embedded on the dashboard of a smart car. Despite the evolution of technology, we have not reached a level where a customer can fully rely on a car; he or she needs to be attentive at all times while using it [87]. The user has the freedom to operate the car using the traditional way, or it can set on auto-pilot temporarily.

Figure 2-36: Tesla

Pal-V

The Pal-V Liberty is another autonomous vehicle, but what makes this one so special is that it can fly as well. The interaction with the user is the most important part, which we will discuss apart from other features. The vehicle is unique because it has two separate engines for flight and another for driving. This vehicle is capable of going to a maximum speed of 100 mph and takes ten minutes to transform into the driving mode or flying mode. When the Pal-V Liberty as shown in Figure 2-37 is in flight mode, it could reach speeds of 112 mph with a maximum range of 817 miles. When the cae is in drive mode, it is 4 meters long, 2 meters wide, and 1.7 meters high. In flight mode, the vehicle measures 6.1 meters long, 2 meters wide, and 3.2 meters high. These vehicles cost between $399,000 to $599,000 and will require a pilot's license to own or drive. Because someone would need a pilot's license to fly, it makes it more complicated not only as a product but also legally. The responsibility and complexity of not only driving but having to fly a vehicle such as this one requires a human to have full focus and attention to operate it, especially when in flight mode. The Pal-V Liberty is made of hand-laid carbon fiber parts, cockpit leather, lightweight aviation aluminum, and an electrical system. It also runs on premium e10 gasoline and gets thity-one miles per gallon in car mode and 6-9 miles per gallon while in the air. The Pal-V liberty was in the works since 2008, with a successful prototype completed in 2009 and the second prototype developed in 2010. It was shown at the Geneva Motor show in March 2018. A basic flying car requires the person at the controls to be both a qualified driver and aircraft pilot. This is impractical for the majority of people, and so wider adoption will require computer systems to simplify piloting. These include aircraft maneuvering, navigation, and emergency procedures, all in potentially crowded airspace. Fly-by-wire computers can also make up for any deficiencies in flight dynamics, such as stability. A practical flying car may need to be a fully autonomous vehicle in which people are present only as passengers [88].

Figure 2-37: Pal-V Liberty

2.3 Discussion

The use of technology is an essential part of our lives. The daily use of devices brings challenges as technology is evolving. There are many instances when systems capable of handling speech input sometimes fail when the user needs technology the most. In such scenarios, the use of highly sophisticated gadgets become a nightmare for the user. There are two broad categories in which users are classified: first, the millennials, for the most part, who are well versed in technology, and second, the senior citizens who are not well versed in technology. The second category finds themselves in huge trouble if provided with technology, and they have to rely on it completely. The use of voice assistants has increased significantly during the last few years, such as Amazon Alexa, Apple's Siri, Google Assistant, and Cortana from Microsoft. The big IT companies are trying their best to overcome the problem recognizing the voice commands inaccurately. Moreover, other modalities such as touch and gesture come with their own set of challenges and a cognitive load to learn the systems. The need of this hour is to design a robus system and capable of managing the input from various users. Here are a few challenges that need to be considered while designing the multimodal system.

2.4 Challenges while designing Multimodal Systems

Designing a multimodal systems is always challenging. The designer should have a broad picture of what kind of requirement they are going to address and who is the user base. Oviatts *Ten Myths of Multimodal Interaction* published in 1999 provide useful insights for those planning to develop a multimodal system.

1. If you develop a multimodal system, the users will interact multimodally. Well-designed systems are set up to allow users to choose their preferred modality for system interaction, which may be unimodal.

2. Multimodal input involves simultaneous signals. Multimodal systems should be designed to allow for sequential use of modalities rather than simultaneous use.

3. Speech and pointing is the dominant multimodal integration pattern.

4. Multimodal integration involves redundancy of content between modes. Usage of varying inputs is preferred over using a single modality again and again.

5. Enhanced efficiency is the main advantage of multimodal systems. However, multimodality does not necessarily increase efficiency; it may or may not. They are designed to provide increased flexibility and increased user satisfaction.

6. Multimodal integration involves redundancy of content between modes.

7. Individual error-prone recognition technologies combine multimodalities to produce even greater unreliability.

8. All users multimodal commands are integrated uniformly.

9. Different input modes can transmit comparable content.

10. Enhanced efficiency is the main advantage of multimodal systems.

2.5 Proposed Multimodal System Design guidelines

Reeves et al. (2004) proposed the guidelines for multimodal systems. Here are the mentioned proposed guidelines [89]:

1. Multimodal systems should be designed while keeping in mind the broadest environment a person could encounter while using the system, for example, use in a private office vs. while driving a car.

2. The designer should consider privacy issues while accessing the system. For example, perhaps a person should be prevented from using speech/voice input while using the system publicly as doing so could lead to a breach of personal information.

3. Maximize human cognitive and physical abilities. The multimodal interface should be designed in such a manner as to be easily understood by the user.

4. Modalities should be integrated in a manner compatible with user preference. For example, users have provided input via speech, they should have the option to receive the output via whatever modality they wish and not just speech. The system should be able to be customize according to user needs.

5. The multimodal system should adapt to the needs and abilities of the user. Individual differences such as age, preferences, pronunciation, sensory skills, and so forth should all be accounted for while designing the system.

6. The output should be consistent and prompt.

7. The system should provide a robust error-handling mechanism.

Table 2.9: Comparison of Multimodal Technologies

Author	Experiment	Modality Used	Multimodal Input	Technology Used	Fusion	Accuracy Claimed	Features	Drawback
Michela	ECOMODE	Haptic	X	Samsung Galaxy S5, iPad mini, and a Samsung Galaxy Tab S 10.5	X	No Empirical Data	Touch screen devices used for Elderly people	The experiments do not yield results in low lighting conditions.
Ian	Baxter Research Robot	Gesture	X	Microsoft Kinect, Baxter robot, PR2 (Kodiak)	X	84 - 89%	X	Detect things which are in front of the robot only.

Author	Experiment	Modality Used	Multimodal Input	Technology Used	Fusion	Accuracy Claimed	Features	Drawback
Natalia	ModDrop	Gesture, Speech	Y	Audio and Video files	Y	96.77%	Multimodal deep learning used for fusing two inputs voice and gesture.	X
Harold	Donaxi	Gesture	X	The robot, and Microsoft Kinect 2	X	2000 iteration required for training	Omni directional navigation system	Extensive training is required before using the system.

Author	Experiment	Modality Used	Multimodal Input	Technology Used	Fusion	Accuracy Claimed	Features	Drawback
Finale	MIT's Human-Robot Interaction	Speech	X	Speech Recognizer (ASR), Natural language (NL) parser, and a dialogue manager (DM)	X	76%	Human speech implemented on a wheelchair	Only accepts speech
Mathieu	CommRob	Speech, Gesture	Y	Java-based FreeTTS API, CommRob Robot	Late Fusion	100% (under restricted conditions)	Speech and Gesture are combined to derive moving of robot at (x,y) coordinates	The experiment did not mention explicitly how they have achieved fusion.

Author	Experiment	Modality Used	Multimodal Input	Technology Used	Fusion	Accuracy Claimed	Features	Drawback
Cristian	Modeling of interaction design	Gesture	X	UI based Modeling tool	X	A satisfactory response in the 4th iteration	X	X
Haleh	Combining Voice and Gesture	Speech, Gesture	Y	Microsoft Kinect API	Y	99.17%	Multimodal Systems proves to be better than Unimodal Systems	Hard to recognize the sound, not reliable and the expeiment can be tiring while using both hands and speech

Author	Experiment	Modality Used	Multimodal Input	Technology Used	Fusion	Accuracy Claimed	Features	Drawback
Nikolas	Natural user interaction	Speech, Gesture	Y	Microsoft Kinect with RGB, depth and audio signal.	Y	No Empirical Data	Proposed architecture could be used with any application running on a computer. Used for moving slides in PowerPoint	Does not have the ability to understand the usage of grammar to understand the input.

Author	Experiment	Modality Used	Multimodal Input	Technology Used	Fusion	Accuracy Claimed	Features	Drawback
Emanuel	MaRDi Dialogue system	Speech, Gesture	Y	3-D simulation software MORSE, KTD-Q algorithms	Y	93.73%	The Pick-Place-Carry task in a Human-Robot Interaction	Interactions take place in a simulated 3-D environment where the user appraisal acquisition is simplified
Emanuel	TownInfo	Speech, Gesture	Y	Simulator using KTD-Q algorithm	Y	95%	A virtual tour guide, better robustness to noisy conditions in terms of semantic input error rate	

Author	**Experiment**	**Modality Used**	**Multimodal Input**	**Technology Used**	**Fusion**	**Accuracy Claimed**	**Features**	**Drawback**
Angelica	Motherese	Speech, Gesture, and input file	Y	Multimodal Emotional Intelligence (MEI) with SIRE input (Speed, Intensity, irRegularity, and Extent).	X	72%	The Motherese robot works with these types of input which are .wav, Vocaloid, Kinect, NAO, Keepon, Flute, Theremin	Accuracy will be improved by training the system in a more diversified environment, e.g., a 20-dimensional confusion matrix.

Author	**Experiment**	**Modality Used**	**Multimodal Input**	**Technology Used**	**Fusion**	**Accuracy Claimed**	**Features**	**Drawback**
Tetsushi	Manipulator Control Interface	Speech, Gesture	Y	7-DOF manipulator (iARM), laptop with a Touch Screen, a CAN-to-USB adaptor, a USB headset	X	No Empirical Data	The robot can easily pick up, rotate and replace objects using gesture and multimodal commands	A user would not be able to operate effectively in a rotational mode
David	Navigation interface	Speech, Gesture	Y	Virtual Geographic Information System (VGIS) and 3-D visualization environment	X	10.1 minutes for 100 commands	3D navigation of the globe.	Users who do not have a clear accent will prefer to use gesture and pen.

Author	Experiment	Modality Used	Multimodal Input	Technology Used	Fusion	Accuracy Claimed	Features	Drawback
Elif	Arm Motion for Prosody-Driven Synthesis	Speech, Gesture	Y	Hidden Markov Model	X	67.80%	Used subjective evaluation methods to set the system parameters and to assess animation quality over two different datasets.	Doesnt include semantic analysis of speech, synthesis of head motion and lip-sync, which would help to achieve more realistic animation results.

Author	Experiment	Modality Used	Multimodal Input	Technology Used	Fusion	Accuracy Claimed	Features	Drawback
Julie	Gesture and voice prototyping	Speech, Gesture	Y	Social acceptability and User perception	X	No Empirical Data	Performance of audible and visible interactions, including how the user perceived the interaction and how comfortable they were while using the device.	Results are not provided in Empirical form. A figure is shown with dots showing the understanding of input from various users.

Author	Experiment	Modality Used	Multimodal Input	Technology Used	Fusion	Accuracy Claimed	Features	Drawback
Nikolas	Multimodal desktop interaction	Speech, Gesture	Y	Microsoft Kinect, Microsoft Speech Recognition SDK.	X	No Empirical Data	1. Login via face detection system, which we have seen recently in Windows 10 2) Application selection via object detection-recognition 3) Authorization control according to log in and data, and 4) Application operations	The research idea is good, but results are collected in terms of successful attempts to open applications on a computer.

Author	Experiment	Modality Used	Multimodal Input	Technology Used	Fusion	Accuracy Claimed	Features	Drawback
Rogalla	ALBERT	Speech, Gesture	Y	ViaVoice	X	95% correct recognition	The experiment uses heuristics of hands by utilizing a webcam available on the robot	Research does not deal with fusion at all, both the experiments being performed separately
Sangseung	Horse Riding Simulators	Speech, Gesture, Haptic	Y	Camera, microphones and other sensors.	X	No Empirical Data	Replicate the mechanistic movements of realistic riding motions	Feature extraction, data fusion, and intention classification is not explained explicitly.

Author	Experiment	Modality Used	Multimodal Input	Technology Used	Fusion	Accuracy Claimed	Features	Drawback
Dennis	Human-robot Interface	Speech, Gesture	Y	PC (Athlon X2 3800), Server programming, Nokia N95 and 40" advanced TV showing the UI	X	No Empirical Data	Works for blind users, visually disabled users, and physically impaired users.	The robot requires specific input speech command to initiate the conversation. Moreover, the robot doesnt work in noisy environments.

Author	**Experiment**	**Modality Used**	**Multimodal Input**	**Technology Used**	**Fusion**	**Accuracy Claimed**	**Features**	**Drawback**
Marjorie	Spatial Language for Human-Robot Dialogs	Speech, Gesture	Y	Personal Digital Assistant, wireless microphone, interacts with a robot via a touch screen and speech.	X	No Empirical Data	The robot will provide detailed spatial descriptions.	Does not facilitate commands concerning objects in the environment, e.g., Move forward until the pillar is behind you.

Author	**Experiment**	**Modality Used**	**Multimodal Input**	**Technology Used**	**Fusion**	**Accuracy Claimed**	**Features**	**Drawback**
Markku	Interface Based on Speech, Gestures, and Haptic Feedback	Speech, Gestures, Haptic	Y	PC, the Athlon X2 3800, Nokia N95, 40 high definition television, and a wireless connector. The application was developed using # and ran under Windows XP.	X	Perceived quality of the speech input surpassed the upper limit of user expectations	Accepts speech input and the response was good.	The interface would work for restricted input.

Author	Experiment	Modality Used	Multimodal Input	Technology Used	Fusion	Accuracy Claimed	Features	Drawback
Stphane	Audio-Visual Speech Modeling	Audio-Visual, Speech	Y	M2VTS audio-visual database	X	95.80%	The database used is extensive with samples collected from 37 different users	The experiment deals with only Speech input. No other input from the user is accepted.
Chieh	Encoder-Camera-Ground Penetrating Radar Tri-Sensor Mapping	Images and GPR/encoder data which are spatially evenly-spaced	Y	Camera, a GPR module which includes control unit, wheel encoder, and GPR antenna.	Y	98%	Developed a encoder-camera-GPR tri-sensor transportation infrastructure inspection sensing suite.	Need to improve further speed and accuracy of the proposed algorithm.

Author	Experiment	Modality Used	Multimodal Input	Technology Used	Fusion	Accuracy Claimed	Features	Drawback
Tom	Road Marking Segmentation via Weakly-Supervised Annotations	Camera captures Images with annotated images.	Y	Camera captured images with annotated images in a weakly-supervised way. Experiment performed on Oxford RobotCar dataset.	Y	75.04%	It has inexpensive manual labelling by exploiting sensor modalities. Useful in creating maps for autonomous vehicles.	The experiment does not include semantic classification of the road markings to retrieve the rules of the road.

Author	**Experiment**	**Modality Used**	**Multimodal Input**	**Technology Used**	**Fusion**	**Accuracy Claimed**	**Features**	**Drawback**
Shan	Vision and Tactile Sensing for Cloth Texture Recognition	Tactile images and vision	Y	GelSight sensor used for capturing camera images and tactile data. Deep Maximum Covariance Analysis (DMCA) algorithm is implemented	Y	90%	Calculated perception performance of either vision or tactile sensing.	Temporal information is not included during the experiments.

Author	Experiment	Modality Used	Multimodal Input	Technology Used	Fusion	Accuracy Claimed	Features	Drawback
Tian	Prediction with Spiking Neural Networks for Human-Robot Collaboration	Gesture, EEG Signals, Speech	Y	MYO armband, Emotiv EPOC, Microsoft Kinect	Y	88%	The experiment exploits unique implementation of Myo armband, Kinect and EPOC Emotiv devices in surgery.	Does not include contextual information to improve early prediction capability e.g., the current status of task progress

Author	**Experiment**	**Modality Used**	**Multimodal Input**	**Technology Used**	**Fusion**	**Accuracy Claimed**	**Features**	**Drawback**
Zhe	Manipulation Graphs from Demonstrations Using Multimodal Sensory Signals	Multimodal Sensory Signals	Y	Barrett arm and hand equipped with two BioTacs	X	80 - 90%	Able to perform grasping, unscrewing, and insertion tasks on a Barrett's arm.	The robotic arm didnt work in those trajectories which are demonstrated earlier.

Author	**Experiment**	**Modality Used**	**Multimodal Input**	**Technology Used**	**Fusion**	**Accuracy Claimed**	**Features**	**Drawback**
Jonas	Multi-Modal Sensor Data for Lower Limb Exoskeletons	Multi-modal sensor data	Y	Hidden Markov Models (HMMs)	X	92.80%	Used for classification of motion patterns at each time step while climbing stairs.	Unable to conduct a deeper analysis of the latencies for different motion transitions.

Author	Experiment	Modality Used	Multimodal Input	Technology Used	Fusion	Accuracy Claimed	Features	Drawback
Castro	Tracking-based Depth Estimation of Metallic Pieces for Robotic Guidance	Image, Lasers	Y	Kinect, Lasers, mono-camera	Y	95%	Performs object recognition and tracking system in real time	The model developed is not used in any application; it just a prototype. To increase the usability, new utilities needs to be incorporated.

Author	Experiment	Modality Used	Multimodal Input	Technology Used	Fusion	Accuracy Claimed	Features	Drawback
Dennis	Multimodal Head-ing and Pointing Gestures for Co-Located Mixed Reality Human-Robot Interaction	Speech and Gesture	Y	Mixed reality inter-face implemented using Microsoft HoloLens	X	93.92%	The interface is capable of guid-ing a robotic arm to picks things	Simple op-eration are performed, unable to in-vestigate if more complex pick poses are requested.

Author	**Experiment**	**Modality Used**	**Multimodal Input**	**Technology Used**	**Fusion**	**Accuracy Claimed**	**Features**	**Drawback**
Zhenzhou	Unsupervised Trajectory Segmentation and Promoting of Multi-Modal Surgical Demonstrations	Video and Kinematic data	Y	Unsupervised deep learning network, stacking convolutional auto-encoder is used.	Y	TSC-K is the biggest beneficiary with the improvement of 15.2% on average	The experiment was designed to accomplish tasks like needle passing and suturing during the surgery.	X

Author	Experiment	Modality Used	Multimodal Input	Technology Used	Fusion	Accuracy Claimed	Features	Drawback
Lus	Extended Bayesian User Model (BUM) for Capturing Cultural Attributes	Sensor simulator and Signal generator	Y	Experiment capture a unified representation of cultural attributes from heterogeneous information.	Y	The framework has a significant impact on classifier precision over time, with an overall improvement of 27.88%.	Ability to learn user attributes from a distributed data stream, with increasing performance over time.	Unable to find information on the user's current state, such as their mood, satisfaction level, etc.

Later in the research, the author discussed multimodal integration and whether it should be performed early or late in the development process. There is no clear-cut answer given to this issue. It varies from system to system. But the authors also presented a big picture which explains what the output of multimodal integration should look like. The system should be able to manage discrete events, loops, and handlers timely and in ways that better match the human interaction the system is intended to support.

2.6 Future Idea for Human-Machine Interaction

As the field is evolving day by day, the researchers are working hard to implement a robust and foolproof solution for daily use. In the era of the Internet of Things (IoT), most of the devices are connected with the network, including household devices such as refrigerators, fans, lights, and even our garages. We believe that User Interface will fundamentally change. The design of an ideal system is not achieved yet, it is because the cost of building a very high precision device is very high. Moreover even the expensive components have probability of error percentage in every device.

1. Hardware equipment is easier to develop and manufacture thanks to innovations like 3-D printers or Arduino.

2. The cost of equipment is dropping significantly on account of the mass adoption of consumer gadgets like cell phones.

3. Technologies like sensors or WiFi chips turn into a generally accessible and cheap commodity and are easy to integrate.

4. As the equipment is becoming easier to develop and manufacture, the focus will move far from innovation and will concentrate more on design or problem solving, the same way it happened in software thanks to the growth of APIs.

5. Software today is built with fifty years of oblivious assumptions of a work tool as a primary concern. There are innovators, particularly from the design world, who get through this presumption and make new UIs [90].

The ideal system designed should consist of the following characteristics:

1. Decentralized: The user interfaces like the light switch shifted onto the smart phone and will now shift away again into smart light switches, speech, or completely new forms like eye tracking.

2. Specific: Interfaces will move far from a nonexclusive screen towards increasingly explicit interfaces that complete only a few things and that are explicitly intended for that utilization case. This implies explicit interfaces for designers that have attention on haptics, interfaces for elderly people that have an emphasis on straightforwardness and unambiguity, or interfaces for children will have an emphasis on playfulness.

3. Human-centered: Graphical UIs have numerous limitations. They are not accessible to visually impaired or disabled people. They utilize the visual sense and a reduced version of haptics. There can be straining to affect our hand, neck or eyes. Future interfaces will be designed with human science and psychology in mind. It will incorporate more of our human senses.

4. Instant: Putting numerous applications on one device implies that the user will need to deal with menus. With decentralized, explicit interfaces, this will be obsolete. Things will be instant again; the question is not whether an action takes 1, 3, or 5 stages. The question will be if an activity should be possible in a split second or not. This additionally diminishes our cognitive load, which enables us to concentrate on the task at hand or the person in front of us.

5. Simple: Future interfaces will disregard the assumed integration with graphical UI and will concentrate on making things less difficult than existing arrangements.

6. Augmented and virtual: The digital and physical will mix together. Whether through augmented reality glasses or not, the user should have the capacity to peruse setting data about a broken device, not through a cell phone but rather specifically in the surrounding ”space” of the object.

7. Passive: An action of a device should be fed as an input to the other. Passive devices are already a major trend in HCI. The classic example of such an application would be turning on the AC of your home when the headlights are approaching the garage [90].

These are few features that a multimodal HCI device must comprise to become successful in the market and accepted by the masses. In the recent past, there are several attempts by companies like which failed deliberately. The Google Glass is one such fine example which became a burden on the user and failed miserably, and the users would not able to leverage the functionalities despite investing a huge chunk of money. Another big factor that also plays a vital role in a device being accepted by a wide range of people is its ”cost”.

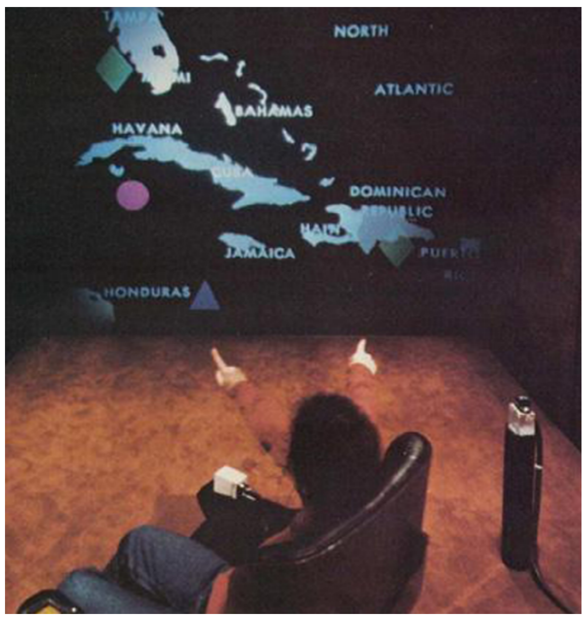

Figure 2-38: Bolts Put That There System (Bolt, 1980).

Chapter 3

Decision Level Multimodal data fusion

3.1 Introduction

Interaction with machines and robots has been a result of the industrial revolution since the last century. The fourth industrial revolution comprises technologies that fuse the physical, digital, and biological worlds, and it affects all disciplines, economies, and industries [91]. It emphasizes designing robots capable of performing tasks while accepting inputs from different modalities such as speech, gestures, and the utilization of peripheral devices. The mode of interaction is selected according to the conditions prevailing in the environment. For example, in a noisy environment of a manufacturing plant, a robot may be operated through gestures and peripheral devices excluding speech or by using a combination of available modalities. Robots are an integral part of our daily lives; they have been widely used in the automotive, health care, manufacturing, and aerospace industries. The manufacturing industry has been shifting to automate processes, as a way of improving quality, reducing human intervention which is more error prone, and, perhaps the largest advantage, relates to the benefits from the lack of fatigue a robot experiences relative to a human worker. This research addresses the challenge faced by users required to control a robot using a set of dedicated controls. In this scenario, we are fusing electromyography data captured from human limbs with human speech. A robot is capable of performing predefined tasks, for example, holding a spare part from a specific location in a 3D environment and inserting it into a machine at a specific location. Previously, this task was performed by robots using a single modality, that is, controlled through a joystick or a programmable board attached to a robot. Now we are trying to make the system more accurate by utilizing a fusion of modalities. The design will incorporate voice commands along with captured muscle movements interpreted via EMG data from the arm-band. The rest of this chpater is organized as follows: Section 2 discusses related work involving multimodal systems with their applications. Section 3, Methodology, addresses both the hardware and the software used, Section 4 provides the results of data fusion based on the experiments performed. Finally, Section 5 discusses various new avenues of research, as well as emerging trends in data fusion using EMG.

3.2 Related Work

New technology often brings with it the idea that machines are not only to be used for the specific predefined purpose, but they can also require the interaction with humans [92]. Human-robot communication is possible through two methods:

- Accepting user input from peripheral devices which are independent of each other, and
- Accepting user input through different modalities and fusing them as a way of obtaining the semantics associated with the actions of the user.

The second approach is the focus of this paper. In 2005, a system was designed that accepts input in the form of speech, keystrokes, and gestures. This system was able to resolve ambiguous inputs and prioritize them [44]. Fusion of multiple inputs is used in several areas of application, and its scope is not only confined to robots, but it also reaches to applications such as authentication systems where fusion could be utilized by, for instance, combining voice recognition and facial detection. A system was designed in 1999 that was able to authenticate a user by comparing inputs against a pre-populated database [93]. The latest version of Microsoft Windows, Windows 10, is capable of authenticating users through a webcam attached to the computer system [94], though it is a unimodal system that could be enhanced with more modalities to improve its accuracy and make it less vulnerable to outside attacks or spoofing. The use of EMG data in fusion is rarely encountered; one such application was implemented to control electronic musical devices through EMG and relative position sensing [95]. The idea of multimodal data fusion has been implemented in industrial robots using the Microsoft Kinect and sensor hardware called Asus Xtion Monitor by capturing hand movements detected by two Leap Motion sensors and performing the resulting mapped actions on a robotic arm [96]. Human-robot interaction during the last five years has largely been performed using the Microsoft Kinect; very few multimodal system designs have used EMG data to fuse with speech, text, and other modalities. The Microsft Kinect is capable of capturing both voice and gestures only on a standalone basis. Moreover, there is no ability to capture EMG data of human limbs using the Kinect. This limitation led us to another niche technology called the MYO sensor arm-band. We have decided to use it as one of the modalities and perform fusion to enhance both its accuracy and performance. The MYO arm-band being an open source software provides avenues to customize gesture and use them in devices used in daily life, for example, controlling a wheel-chair, turning a door knob, and so on [97]. The results achieved with other experiments to evaluate the accuracy of MYO comes out to 87.8 to 89.38 percent which provides us avenues for improvement [98]. The MYO band is used in an experiment both to perform searches and select operations on a computer, and the average score for evaluation is analyzed. The researchers conclude that after adapting the limbs temperature the MYO band constantly performs with a similar number of scores [99]. It is our belief that is this the first research work undertaken that uses EMG data to capture gestures using MYO armband sensors for multimodal data fusion.

3.3 Methodology

The robots first introduced to the market were relatively simple most of them required a teaching phase and programming. More recently, robots have become dynamic, sophisticated, and capable of much more than before [100]. Along with this sophistication came increasing demands to perform complex tasks which require both accuracy and precision. The standalone robot in the experiment introduced in this research showed ample room for improvement in both robustness and accuracy. Hence it was decided to improve the accuracy of a robotic arm through the use of multi-modal data fusion. The experiment emphasizes the conversion of input data through different channels into a single format, which is understood by the robotic arm through mediation. The input modalities used are speech and gesture.

3.3.1 Hardware

The system designed for the experiment described in this research is composed of the following components: an Arduino based robotic arm and a MYO armband. These devices are illustrated in Figure 3-1 and Figure 3-2 respectively. The robotic arm used is manufactured by Trossen Robotics. The robot used in the experiment described here is called RobotGeek Snapper Arduino Robotic Arm, and it contains five servo motors. The robotic arm is controlled by an electromyography data-based armband called a MYO. Figure 3-1 depicts the usage of the MYO band from which data is captured and manipulated to perform actions on various Arduino-based devices. The armband is capable of capturing five gestures: fist, wave left, wave right, double tap, and fingers spread, as shown in Figure 3. This arm band provides the ability to customize an open library and perform actions according to the user's need. The robotic arm used comprises Arduino Duemilanove and Diecimila boards for accepting input through USB. The Arduino board is connected to the robot with pins defined for each specific motor, as shown in Figure 3-2. A high-precision wireless H800 headset from Logitech is used for capturing the speech input.

Figure 3-1: MYO Arm Band

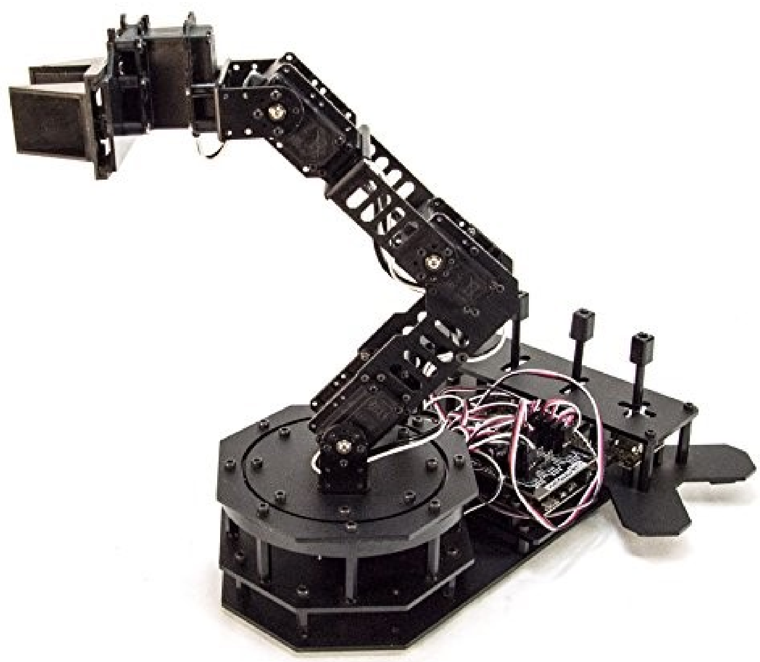

Figure 3-2: Arduino based robotic arm

3.3.2 Software

The software is implemented using C# and C++ programming languages. C++ is used to implement the MYO API while C is used for providing speech input using Microsofts speech engine. Initially, the robotic arm was operated using gestures only. The MYO armband was used to support the communication with the robot. It was concluded from experimental results that the precision of the armband was not very high. To improve the accuracy of the commands communicated fusion of the input data was introduced. The speech modality was then combined with the set of gestures previously defined. The MYO band provides an API to control the Arduino board, which is connected to the robotic arm, as shown in Figure 3-2. The Arduino board consists of fourteen pins and provides the capability to connect each servo motor of the robotic arm.

3.3.3 Experimental Setup

The experimental system is designed using a client-server paradigm. The MYO armband consists of eight sensors, which capture the muscle movement. Such sensors compare and match hand movements to gestures defined in MYO. For example, the Wave Out gesture listed in Table 3.2 is performed by moving the hand in a vertical orientation toward the right, as shown in the third gesture of Figure 3-3. The MYO program is functioning as the server while the Microsoft speech program is designed as the client. As mentioned above, the accuracy of capturing gestures is not very high, which tends to result in matching the hand movement to the wrong gesture, for instance, Wave Out may be captured as Wave In. Moreover, the band sometimes fails to capture a gesture entirely. Both of these cases are considered errors. The MYO API allows it to be customized according to the project needs, and the prototype has thus implemented threads responsible for listening to gestures. If a gesture is missed by the band, voice commands compensate for the missed input through human speech. Processing of speech input is implemented in a client component, which sends commands to the server (MYO API). Priority is given to the MYO band, but in the case of an error, speech recognition activates and helps in improving the accuracy of controlling the robot. Fusion is thus performed in the order of priority. Gestures are given the highest priority. In case of a failure to capture the input, voice commands are used to compensate and serve as the only input. Priority-based fusion is used in other domains, including medical systems, and tends to improve its accuracy significantly [101].

The speech is fused with EMG input received from MYO, which enables the robot to work precisely according to the user's input command. When the user performs gestures using his or her arm, the input message is transmitted from the MYO band to the Arduino, and as a result, it moves the specific servo motor. The fused input sets

Figure 3-3: Gestures available with MYO band

the corresponding Arduino pin to high, namely, 1, which then moves the robot. In the prototype constructed, five different Arduino pins were linked to various gestures, as shown in Table 3.1. Through the combination of gestures and speech, users should be able to control the robotic arm precisely and accurately. The results are shown in the next section.

Table 3.1: Mapping of Gestures with Arduino Boards

Fist	Pin 3
Wave In	Pin 4
Wave Out	Pin 5
Finger Spread	Pin 9
Double Tap	Pin 10

Table 3.2: Preliminary results for of Muscle sensor MYO band

Gesture	**Wrong/Missed gesture %**	**Correct %**
Wave Out	9.5	90.5
Wave In	9.1	90.9
Fist	13.6	86.4
Double Tap	20.6	79.4
Finger Spread	14.5	85.5

3.4 Results and Discussion

A performance evaluation was executed to quantify how accurate the modalities are individually, and thus we tested them separately. The Microsoft Speech API was tested using the simple speech commands such as, for example, move right, move left and so on. The complete list of speech commands is illustrated in Table 3.3. The experimental results show the scope of improvement as the error for the speech API lies between 8.9 and 34.2 percent. Similarly, the MYO band results were captured to quantify accuracy and to find the scope of improvement [6].. The preliminary results have shown the MYO band has a scope of improvement. The error rate lies between 9.1 and 20.6 percent. The data has been collected by experimenting ten times, with each experiment having one hundered gestures performed and then calculating the average percentages shown in 3.2. An error for the experiment occurs when a gesture is either missed or captured wrong. The trials have been performed in laboratory conditions. The MYO armband is capable of adapting to specific human limbs and improves its output once it has been trained completely. The arm band sensors become warm up shortly after being put on, thus adapting to body temperature and accurately recognizing gestures after one to two minutes. If one's arm is cold, the sensors are unable to capture gestures accurately. The Microsoft API results are evaluated by speaking a command, and if the command is captured incorrectly, the instance is marked as an error. Incorrect capture is defined as the resulting string being captured twice or having extraneous words or characters added to it. Multimodal data fusion of voice and gesture using the MYO band improves the system performance significantly. The experimental results are shown in Table 3.4. After implementing fusion in the robotic arm, the error rate is reduced to 5.2 percent which is an average of all errors. The variance of error percentage is shown in 3-4. The errors are mostly due to reading the wrong gesture, for example, finger spread is sometimes captured as a fist, which leads to an error. Experiments are performed on all five fusion input tests two hundreds times each, and the percentage is calculated respectively.

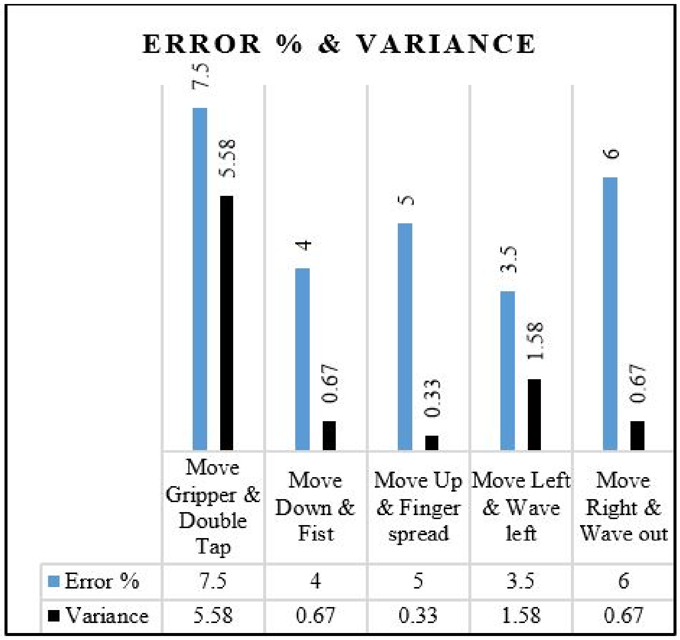

Figure 3-4: Error Percentage deviation of fused Inputs

Table 3.3: Preliminary results for Microsoft Speech used by non-native speaker.

Command	Microsoft Speech API (Wrong output)	Correct
Move Right	10	90
Move Left	34.2	65.8
Move Up	8.9	91.1
Move Down	22.5	77.5
Move Gripper	14.1	85.9

Table 3.4: Fusion results with Error % & Variance

Fusion Operation	**50**	**100**	**150**	**200**	**Error %**	**Variance**
Move Gripper & Double Tap	7	2	2	4	7.5	5.58
Move Down & Fist	3	1	2	2	4	0.67
Move Up & Finger spread	3	3	2	2	5	0.33
Move Left & Wave left	0	3	2	2	3.5	1.58
Move Right & Wave out	3	3	4	2	6	0.67

Chapter 4

Machine Learning based Multi-modal data fusion

To design a system with multiple modalities, bring its own challenges and complexities. The system design with multiple input methods requires a mechanism to take a decision. There are several approaches available, which include the implementation of machine learning and Artificial Intelligence. Machine learning itself is divided into two broad categories, namely supervised learning and unsupervised learning. Supervised learning learns and decides which decision the system should make to work efficiently. It learns based on data provided initially, and then it learns gradually while the system is in use. The other approach is unsupervised learning based on the categorical classification of huge data; it helps in sorting the news into several categories, the classification of email in various labels, and so on. In our case, we are planning to implement supervised learning. The idea to be implemented in our case is to make the robotic arm capable of learning from its previous decisions. Based on the previous history, the system would make a decision and improve the decision.

4.1 Machine Learning

Machine Learning has four broad categories: : unsupervised learning for clustering, unsupervised learning for dimension reduction, supervised learning for classification, and supervised learning for regression.

1. Unsupervised learning for clustering of data - this approach is used when the bulk data provided without any response. It helps in clustering in several scenarios. For classification of data on the based-on hierarchy, an algorithm named as hierarchical machine learning algorithm is used. DBScan is another algorithm used, in which we dont have defined levels of hierarchy. The data provided for unsupervised learning can be classified based on probability; the Gaussian Mixture Model is used for finding the probability from a data set. Moreover, K-means and K-modes are used for categorical classification of data [102].

2. Unsupervised learning for dimension reduction - this approach is used when the intent is to reduce the number of random variables under consideration by obtaining a set of principal variables. The algorithm called Principal component analysis is used for reducing the number of random variables, while singular value decomposition and latent Dirichlet analysis for classifying data based on probability [102].

3. Supervised learning and classification of data is a classification model that attempts to draw conclusions from observed values. One or more input is provided to the classification model, and it will try to predict the value with one or more outcomes. The system is trained first with inputs, and accordingly, later it takes actions. There are several algorithms available for supervised learning based on speed, accuracy, and size of data. Kernel SM, Random Forest, gradient boosting tree are the algorithms available for accuracy, while Nave Bayes is used when

data is too large. Decision tree and logistic regression are used for processing data quickly [102].

4. Supervised learning for regression is used for numeric predicting regarding 0 or 1. For accuracy, the algorithms available are Random Forest, Gradient Boosting tree, and Neural Network. For speed, decision tree and linear regression are used [102].

These machine learning models which are available & implemented in Python:

1. Logistic Regression
2. Linear Discriminant Analysis
3. K-Neighbors Classifier
4. Decision Tree Classifier
5. Gaussian Nave Bayes
6. Support Vector Machine

Logistic Regression: Logistic regression is a machine learning technique extracted from the field of statistics. It is the go-to method for binary classification problems (problems with two class values). Logistic regression named for the function used at the core of the method, the logistic function as shown in Figure 4-1. The logistic function, also called the sigmoid function, was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It is a S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits [103].

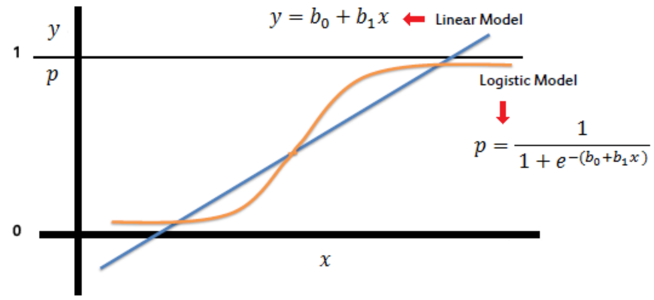

Figure 4-1: Data Mining graph of Logistic Regression

Linear Discriminant Analysis: Linear discriminant analysis (LDA) is commonly used as a dimensionality reduction technique in the pre-processing step for pattern-classification and machine learning applications. The goal is to project a dataset onto a lower-dimensional space with good class-separability to avoid overfitting (curse of dimensionality) and reduce computational costs [104].

The K-neighbors classifier has been used widely in pattern recognition. The algorithm is easy to understand conceptually, and the tendency toward error is bounded twice by the Bayes error. The accuracy of K-neighbor surpasses those of sophisticated classifiers. The random subspace method relies on a stochastic process that randomly selects components [105]. K-nearest neighbors algorithm (KNN) is a non-parametric, lazy learning algorithm. Its motivation is to utilize a database in which the information focuses are isolated into a few classes to anticipate the characterization of another sample point, as shown in Figure 4-2.

Decision Tree Classifier: Decision trees are a type of supervised machine learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely, decision nodes and leaves. The leaves are the decisions or the outcomes, and the decision nodes are where the data is split. An

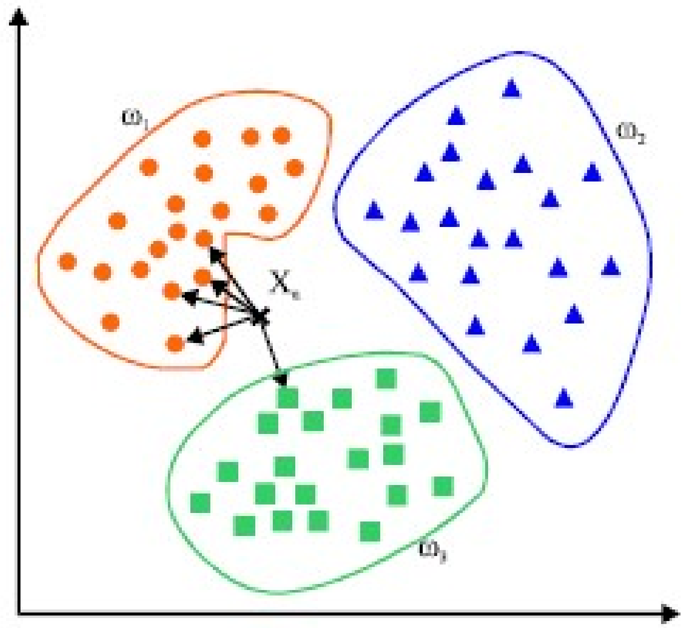

Figure 4-2: K-NN classifier: Classifying using k-nearest neighbors algorithm

example of a decision tree can be explained using the above binary tree. Let's say you want to predict whether a person is fit based on his or her information like age, eating habits, physical activity, and so on. The decision nodes here are questions like 'What's the age?', 'Does he exercise?', and 'Does he eat a lot of pizzas'? And the leaves are outcomes like fit, or unfit. In this case, this was a binary classification problem (a yes/no type problem) [106].

Gaussian Naive Bayes: Gaussian naive bayes is a classification algorithm for binary (two-class) and multi-class classification problems. The technique is easiest to understand when described using binary or categorical input values. The approach is called naive Bayes or idiot Bayes because the calculation of the probabilities for each hypothesis is simplified to make their calculation tractable. Rather than attempting to calculate the values of each attribute value P(d1, d2, d3—h), they are assumed to be conditionally independent given the target value and calculated as P(d1—h) * P(d2—H), and so on. This is a very strong assumption that is most unlikely in real data, that is, that the attributes do not interact. Nevertheless, the approach performs surprisingly well on data where this assumption does not hold [107].

Support Vector Machine: Support vector machine (SVM) is a supervised machine learning algorithm used for both classification or regression challenges. However, this is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is the number of features present) with the value of each feature being the value of a coordinate. Then we perform classification by finding the hyperplane that differentiates the two classes [108].

Policy-gradient methods are reinforcement learning techniques that rely on optimizing parametrized policies concerning the expected return (long-term cumulative reward) by gradient descent. Policy gradient algorithms optimize a policy by computing the gradient of the expected reward of the policy and then updating the policy in the gradient direction. A stochastic policy is preferred as it gives a randomized probability distribution over actions. The algorithms requires many training examples designed so that good actions result in high rewards while bad actions result in negative rewards. The observations can then be used to increase the probability that the algorithm chooses from the set of good actions. Most of the problems in the reinforcement learning space involve a single reward signal generated at the end of an actor-environment simulation episode. It makes it difficult to identify the good actions from the set of all actions taken during that episode, and this is known as the credit assignment problem. The problem is more evident in cases where the action spaces are continuous, and actor-critic methods are used to solve this problem for continuous action spaces [109].

Actor-critic methods are TD (temporal-difference) methods. In this algorithm, the policy function is represented independently by the value function. The policy structure is known as the actor because it is used to select actions, and the estimated value function is known as the critic because it criticizes the actions made by the actor. It is an on-policy method because the critic must learn and critique the policy that is currently being followed by the actor. The actor produces an action depending on the current state of the environment, and the critic produces a TD error signal depending on the state and resultant reward. To train an actor-critic algorithm, the initial state is observed, and the actor chooses an action from the set of all available actions and then observes the resultant state. After the critic network has assigned a value to both the original and the new state based on a reward function, we adjust the policy. This is achieved by comparing the value function in the new state and the original state; if the value improves, we encourage that action. If it decreases, we discourage the action [110]. During initial training, both networks generate a lot of bad choices. In deep reinforcement learning, neural networks can be used to represent the actor and critic structures.

Deep deterministic policy gradient (DDPG) Algorithm [111] is a policy gradient algorithm that uses a stochastic behavior policy to reduce the predictability of the learned model but estimates a deterministic target policy, which is much easier to learn. Stochastic behaviors are the situations or models containing a random element. Hence they are unpredictable and without a stable pattern or order. All-natural events are considered stochastic phenomena [112].

A stochastic behavior policy performs better for problem domains where exploratory actions are required for reaching the solution and helps prevent against convergence to a local minimum. DDPG is an off-policy algorithm and uses a deterministic target policy, which allows for the use of the deterministic policy gradient theorem as proven by Silver et al. (2014). The ability of DDPG to operate over continuous action space makes it suitable for use in our work. Twelve DDPG algorithm Q-learning cannot be applied directly for continuous action space. This is because finding the greedy policy in continuous spaces requires optimization of action at every time step; this optimization is not practical because of the large non-trivial action spaces. DDPG achieves this using the actor-critic approach. The actor-critic function helps to represent the policy function independent of the value function. The actor takes as input the current state of the environment and gives an action as an output. The critic gives a temporal difference error signal based upon the state and the resultant reward. The output obtained from the critic is used to update both the actor and critic. The actor and critic structures are modeled as neural networks that try to choose an action from the continuous action space according to the current state to try to minimize the TD error signal at each time step. However, when using neural networks for reinforcement learning, the algorithm assumes that the input samples are independent and identically distributed. However, this assumption is wrong as the inputs obtained are sequential. Tackling this DDPG requires a finite-sized buffer representing historical states called a replay buffer first proposed by Timothy Lillicrap et al., (2015) [111]. All inputs to the actor are sampled from a minibatch from the replay buffer. Once the replay buffer is full, the oldest samples are removed. The input to the actor network is the current state, and the output is a single real value representing an action chosen from a continuous action space. The critic outputs the estimated Q-value of the current state and the action chosen by the actor. The actor is updated using the deterministic policy gradient theorem. The critic is updated from the gradients obtained from the TD error signal. DDPG also makes use of batch normalization [113] to normalize each dimension across the samples to have unit mean and variance. This helps to address the issue that different components of the observation vector may have different physical units such as distance, velocity or acceleration. Batch normalization helps DDPG to learn across different units in its observation vector. In order to treat the problem of exploration of the continuous action space as an independent problem, which can be modeled using a noise process to assist exploration using the actor policy, the Ornstein-Uhlenbeck [114] process is used to generate temporally correlated noise and is particularly suited for problems involving physical control.

4.2 Feature Level and Decision Level Fusion

Multimodal fusion is the heart of any multimodal sentiment analysis engine. There are two main fusion techniques: feature-level fusion and decision-level fusion. Feature-level fusion is implemented by concatenating the feature vectors of all three modalities to form a single long feature vector. Despite its simplicity, this method produces accurate results. We concatenated the feature vector of each modality into a single feature vector stream. This feature vector is used for classifying each video segment into sentiment classes. To estimate the accuracy, we used tenfold cross-validation.

In decision-level fusion, we obtained feature vectors instead of concatenating the feature vectors as in feature-level fusion, we used a separate classifier for each modality. The output of each classifier is treated as a classification score. We obtained a probability score for each sentiment class, from each classifier. In our case, as there are three sentiment classes, we obtained three probability scores from each modality.

4.3 Result

A performance evaluation was executed, keeping in mind the end goal to evaluate how exact the modalities are independently; subsequently, we tried them using machine learning techniques. The figure 4-4 shows the data density. Based on the data retrieved in Phase 2, training data is prepared for machine learning algorithms. Both the modalities will be represented by a number and third column as an output. Input data of 900 interactions will be fed to the machine learning algorithm, test results of each modality and its error cases are considered. The error inputs are considered with a numeric value; for instance, Move Up is could be read as Move Cup or Move Sup by Microsoft Speech API. If there are 59 error combinations, the test data for 59 error conditions are created, and in case of an error, the other correct modality if given the priority.

Similarly for EMG data, if the Myo armband captures the wrong gesture or misses the input, both the scenarios are considered an error. The conclusion is evaluated based on the comparison of various Machine learning algorithms on the same training data. Table 4.1 shows that K-NN has the highest precision and recall with the value of 0.92. The F-1 score for K-NN is 0.92, and the support is 1, which is highest amongst all the algorithms. The 4-4 shows data density of EMG. Speech and the Output. The complete details of the the results are outlined in Table 4.2. The test results demonstrate the best algorithm best work for us is K-Neighbors Classifier with an accuracy of 92.45 percent.

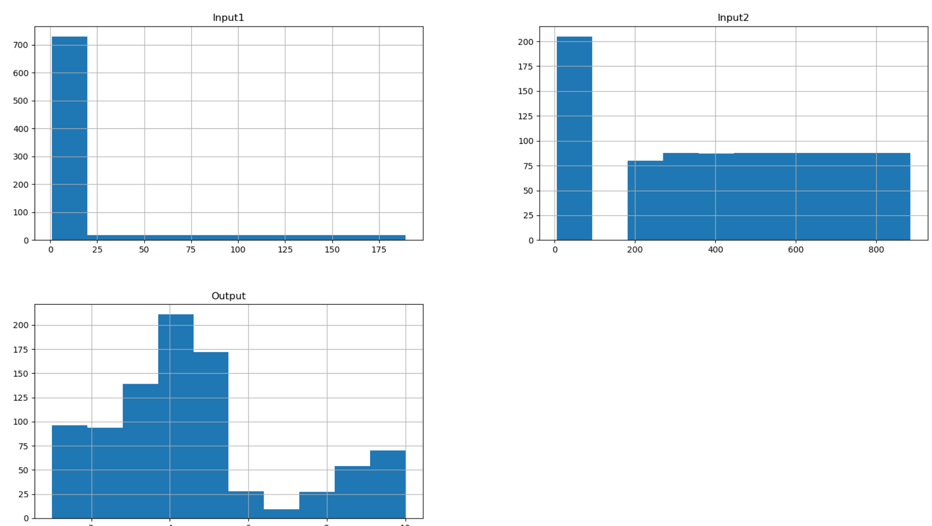

Figure 4-3: Input Data

4.4 Limitations

Speech and electromyography data were the modalities used in the system constructed. The MYO band used for capturing the EMG data is capable of recognizing

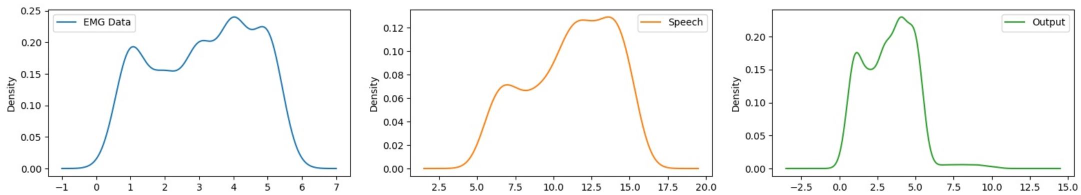

Figure 4-4: Data Density

Table 4.1: Precision, Recall and F1-Score

Machine Learning Algorithm	Precision	Recall	F1-Score	Support
SVM	0.82	1.00	0.90	9
Gaussian NB	0.67	0.50	0.57	4
Decision Tree Classifier	0.83	0.83	0.83	6
Linear Discriminant Analysis	1.00	0.75	0.86	8
Logistic Regression	0.78	1.00	0.88	7
KNN	0.92	0.92	0.91	1

five gestures. This limited the number of operations that could be performent on the robotic arm. The second challenge lay in capturing the speech commands using the Microsoft Speech API. Non-native speakers of the English language will face difficulties and challenges to approximate their accent to that of a native speakers. This created difficulty in conveying commands correctly. There are four areas that need to be worked on and improved regarding human-robot interaction. These include speech localization, language understanding, dialogue management, and speech synthesis [115]. Also, as the ultimate goal of this research is to improve accuracy, the approach here described map commands to all possible options that the Microsoft Speech API recognizes as valid (for example, move right sometimes gets recognized as override an incorrect response). This provided us with a way to quantify the accuracy of the system. We prepared a many-to-one mapping of all these possible combinations to a particular voice command.

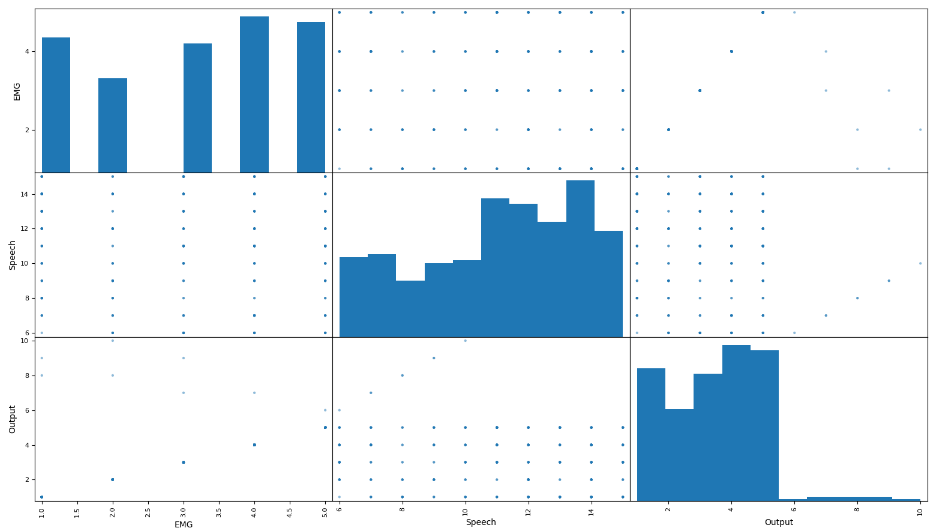

Figure 4-5: Scatter Plot

Table 4.2: Results of Machine learning algorithms implemented on Multi-modal input

Machine Learning Technique	**Correct**
Logistic Regression	84.90
Linear Discriminant Analysis	84.90
K-Neighbors Classifier	92.45
Decision Tree Classifier	84.90
Gaussian Naive Bayes	86.79
Support-Vector Machine	90.56

Chapter 5

Conclusion and Future Work

With the classification of multimodal systems, there is an ample amount of space for improvement regarding precision, quality, and robustness. The implementation of the multimodal system using supervised learning has high precision and can capture various modalities. The system, designed recently in 2018 and 2019 has advanced in exploring different modalities and its applications, including their application from surgery to find the texture of the cloth and fixing the screw using the robotic arm. This study enables us to develop a system that can implement multimodal fusion using voice, gesture and peripheral device as an input. In most of the systems with fusion, we have analyzed has a huge scope of improvement. In our future work, we plan to develop a system that captures raw EMG data using human limbs along with voice, including input from a third modality, which could be a lever or a keyboard input, that allows to generate input by combination of modalities if the robot misses a gesture or a speech. The initial phase has been implemented and incorporates two modalities, speech and gesture [6]. The results show significant improvement in comparison with individual modality, with the average error rate reduced to 5.2 percent. Age plays an important factor when using multimodality. Younger people are more comfortable with MYO arm band and other peripheral devices devices , whereas older adults prefer devices with key input in comparison to haptic and gestures.

Moreover, the voice input modality is still preferable, but it brings another challenge in that it does not accept all the accents of various ethnicities, and the shaky voices of older adults. An incorrect input of voice commands may lead to disaster in situations involving driverless cars. A multimodal autonomous car should accept canned pre-defined input and discard others to function properly and seek minimum human intervention. Otherwise, if a car analyzes human speech and receives wrong input, it could be fatal for the users. Likewise, giant robots used in automation or manufacturing units, if developed with multimodal functionality, must understand canned inputs and discard others to function properly. The assisted robot for the blind project is one of the cutting-edge projects in the field of multimodal inputs, but if the system is not efficient enough to adapt, the environment will fail during the evacuation of a building when developing systems for differently abled people, efficiency and accuracy would be considered the most important criteria. Otherwise, this robotic system might prove fatal to humans. There is a need to develop a device of multimodality with fusion that can be used in myriad industries. During the review, we did not come across a system that can accept multimodal inputs with fusion and is efficient enough to perform tasks in an industrial or a health care sector. So far, the systems designed are either multimodal without fusion, or they accept pre-defined inputs that work under certain conditions.

Moreover, in an industrial environment, it is necessary to move a robot dynamically in any direction to move heavy objects from one place to another. Usually, assembly lines can perform predefined static tasks. Along the similar lines, a speech assistant with an intelligent robotic arm is needed to understand human speech irrespective of ethnicity and pronunciation and that can guide a differently abled person to move in and around a city without the help of other human beings. Google and Microsoft developed speech APIs, which are paid and too expensive to be used by everyday people, and they also have the challenge of understanding of human speech with high accuracy. An error in understanding speech could lead to disaster and may endanger human lives. The ECOMODE device discussed earlier brings a lot of challenges to be used by senior citizens. Emotions and gestures are an easy way to capture input by avoiding speech.

According to our study, there are still avenues open for research in multimodal fusion while exploring different combinations of input modalities. If a system is designed along similar lines, it not only improves the capability of handling industrial robots but also makes life easier for the differently abled.

The experimental results displayed above prove that through the inclusion of speech input modality, the accuracy of the MYO band can be improved significantly. While using the modalities separately, the accuracy was 86.54 percent and 82.06 percent for the MYO band and Microsoft Speech API, respectively. After fusion of the inputs, accuracy improved to more than 95.92 percent. Our future work will include the development of a prototype in which the system can perform fusion of more than two input modalities and perform tasks after interpreting the semantics of the input provided. Thus far there is no system that takes input from the user in the form of speech, text, and gesture and executes a task on a robotic arm using the MYO band dynamically. The next planned implementation will add components capable of capturing brain signals. The system should be able to fuse the modalities and select a meaningful operation that is then performed on a device.

References

[1] P. Adkar, "Unimodal and multimodal human computer interaction: a modern overview," *Pratibha-International J. Comput. Sci. Inf. Engg*, vol. 2, pp. 2277–4408, 2013.

[2] S. Oviatt, "Ten myths of multimodal interaction," *Communications of the ACM*, vol. 42, no. 11, pp. 74–81, 1999.

[3] D. Lalanne, L. Nigay, P. Robinson, J. Vanderdonckt, J.-F. Ladry, *et al.*, "Fusion engines for multimodal input: a survey," in *Proceedings of the 2009 international conference on Multimodal interfaces*, pp. 153–160, ACM, 2009.

[4] S. Keates and P. Robinson, "The use of gestures in multimodal input," 1998.

[5] L. Zhang, J.-Q. Li, K. Zhou, S. D. Gupta, M. Li, W.-B. Zhang, M. A. Miller, and J. A. Misener, "Traveler information tool with integrated real-time transit information and multimodal trip planning: Design and implementation," *Transportation Research Record*, vol. 2215, no. 1, pp. 1–10, 2011.

[6] T. K. Mohd, J. Carvalho, and A. Y. Javaid, "Multi-modal data fusion of voice and emg data for robotic control," in *2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*, pp. 329–333, IEEE, 2017.

[7] S. Vyamajala, T. K. Mohd, and A. Javaid, "A real-world implementation of sql injection attack using open source tools for enhanced cybersecurity learning," in

2018 IEEE International Conference on Electro/Information Technology (EIT), pp. 0198–0202, May 2018.

[8] Y. Kolli, T. K. Mohd, and A. Y. Javaid, "Remote desktop backdoor implementation with reverse tcp payload using open source tools for instructional use," in *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pp. 444–450, Nov 2018.

[9] F. Karray, M. Alemzadeh, J. A. Saleh, and M. N. Arab, "Human-computer interaction: Overview on state of the art," 2008.

[10] N. Mavridis, "A review of verbal and non-verbal human–robot interactive communication," *Robotics and Autonomous Systems*, vol. 63, pp. 22–35, 2015.

[11] J. C. Augusto, H. Nakashima, and H. Aghajan, "Ambient intelligence and smart environments: A state of the art," in *Handbook of ambient intelligence and smart environments*, pp. 3–31, Springer, 2010.

[12] T. Carter, S. A. Seah, B. Long, B. Drinkwater, and S. Subramanian, "Ultrahaptics: multi-point mid-air haptic feedback for touch surfaces," in *Proceedings of the 26th annual ACM symposium on User interface software and technology*, pp. 505–514, ACM, 2013.

[13] M. Ferron, N. Mana, and O. Mich, "Mobile for older adults: towards designing multimodal interaction," in *Proceedings of the 14th International Conference on Mobile and Ubiquitous Multimedia*, pp. 373–378, ACM, 2015.

[14] "Boschs view of the future car: truly keyless entry, haptic feedback, smart parking, cybersecurity." `https://bit.ly/2Y9eVe6`. (Accessed on 06/06/2019).

[15] S. Mitra and T. Acharya, "Gesture recognition: A survey," *IEEE Transactions*

on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 37, no. 3, pp. 311–324, 2007.

[16] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *The International Journal of Robotics Research*, vol. 34, no. 4-5, pp. 705–724, 2015.

[17] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, "Moddrop: adaptive multi-modal gesture recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1692–1706, 2015.

[18] H. A. Vasquez, H. S. Vargas, and L. E. Sucar, "Using gestures to interact with a service robot using kinect 2," *Advances in Pattern Recognition*, vol. 85, 2015.

[19] "How does the leap motion controller work?." `http://blog.leapmotion.com/hardware-to-software-how-does-the-leap-motion-controller-work/`. (Accessed on 06/07/2019).

[20] "How does the leap motion controller work?." `http://blog.leapmotion.com/`. (Accessed on 06/11/2019).

[21] "Getting started with your myo armband." `https://support.getmyo.com/hc/en-us/articles/203398347-Getting-started-with-your-Myo-armband`. (Accessed on 06/12/2019).

[22] C. M. Rebman Jr, M. W. Aiken, and C. G. Cegielski, "Speech recognition in the human–computer interface," *Information & Management*, vol. 40, no. 6, pp. 509–519, 2003.

[23] F. Doshi and N. Roy, "Spoken language interaction with model uncertainty: an adaptive human–robot interaction system," *Connection Science*, vol. 20, no. 4, pp. 299–318, 2008.

[24] "Human-computer interaction: Why voice changes everything." https://medium.com/@joshdotai/human-computer-interaction-why-voice-changes-everything-f16cdc9df39. (Accessed on 06/09/2019).

[25] "Siri." https://en.wikipedia.org/wiki/Siri. (Accessed on 06/06/2019).

[26] "Cortana. your intelligent assistant across your life.." https://www.microsoft.com/en-us/cortana. (Accessed on 06/07/2019).

[27] "The eyegaze edge." https://eyegaze.com/products/eyegaze-edge/. (Accessed on 06/06/2019).

[28] "Immersive reality without a headset." https://gaming.tobii.com/product/tobii-eye-tracker-4c/. (Accessed on 06/25/2019).

[29] "Okay, but how do touch screens actually work?." https://scienceline.org/2012/01/okay-but-how-do-touch-screens-actually-work/. (Accessed on 06/08/2019).

[30] "Touchscreen." https://en.wikipedia.org/wiki/Touchscreen. (Accessed on 06/06/2019).

[31] "Galaxy fold." https://www.samsung.com/global/galaxy/galaxy-fold/design/. (Accessed on 06/07/2019).

[32] "Meet nintendo switch." https://www.nintendo.com/switch/features/. (Accessed on 06/07/2019).

[33] "Digital pens." https://www.explainthatstuff.com/digitalpens.html. (Accessed on 06/07/2019).

[34] "Field mapping for agriculture." `https://www.skyswoop.com/services/field-mapping-for-agriculture/`. (Accessed on 06/07/2019).

[35] "Epoc user manual." `https://emotiv.gitbook.io/epoc-user-manual/introduction-1/about`. (Accessed on 06/12/2019).

[36] S. Oviatt, "Multimodal interfaces," *The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications*, vol. 14, pp. 286–304, 2003.

[37] R. A. Bolt, *Put-that-there: Voice and gesture at the graphics interface*, vol. 14. ACM, 1980.

[38] T. Zaheer, "Multimodal user interfaces for newspaper distribution transcription on mobile devices with accelerometer sensors," Master's thesis, 2009.

[39] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia systems*, vol. 16, no. 6, pp. 345–379, 2010.

[40] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: an overview of methods, challenges, and prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.

[41] Y. Vershinin, "A data fusion algorithm for multisensor systems," in *Proceedings of the Fifth International Conference on Information Fusion. FUSION 2002.(IEEE Cat. No. 02EX5997)*, vol. 1, pp. 341–345, IEEE, 2002.

[42] V. Pitsikalis, A. Katsamanis, S. Theodorakis, and P. Maragos, "Multimodal gesture recognition via multiple hypotheses rescoring," *The Journal of Machine Learning Research*, vol. 16, no. 1, pp. 255–284, 2015.

[43] J. Wagner, F. Lingenfelser, T. Baur, I. Damian, F. Kistler, and E. André, "The social signal interpretation (ssi) framework: multimodal signal processing and recognition in real-time," in *Proceedings of the 21st ACM international conference on Multimedia*, pp. 831–834, ACM, 2013.

[44] M. Turk, "Multimodal interaction: A review," *Pattern Recognition Letters*, vol. 36, pp. 189–195, 2014.

[45] D. M. Krum, O. Omoteso, W. Ribarsky, T. Starner, and L. F. Hodges, "Speech and gesture multimodal control of a whole earth 3d visualization environment," tech. rep., Georgia Institute of Technology, 2002.

[46] M. Vallee, B. Burger, D. Ertl, F. Lerasle, and J. Falb, "Improving user interfaces of interactive robots with multimodality," in *2009 International Conference on Advanced Robotics*, pp. 1–6, IEEE, 2009.

[47] C. Bogdan, H. Kaindl, J. Falb, and R. Popp, "Modeling of interaction design by end users through discourse modeling," in *Proceedings of the 13th international conference on Intelligent user interfaces*, pp. 305–308, ACM, 2008.

[48] A. Kendon, *Gesture: Visible action as utterance.* Cambridge University Press, 2004.

[49] H. Chizari, D. Lalanne, and M. Schwaller, "Combining voice and gesture for human computer interaction," *Master's thesis, Freiburg University*, 2013.

[50] N. Vidakis, M. Syntychakis, G. Triantafyllidis, and D. Akoumianakis, "Multimodal natural user interaction for multiple applications: The gesturevoice example," in *2012 International Conference on Telecommunications and Multimedia (TEMU)*, pp. 208–213, IEEE, 2012.

[51] E. Ferreira and F. Lefevre, "Reinforcement-learning based dialogue system for human–robot interactions with socially-inspired rewards," *Computer Speech & Language*, vol. 34, no. 1, pp. 256–274, 2015.

[52] V. Vielzeuf, A. Lechervy, S. Pateux, and F. Jurie, "Multilevel sensor fusion with deep learning," *IEEE sensors letters*, vol. 3, no. 1, pp. 1–4, 2019.

[53] M. S. Hossain and G. Muhammad, "Emotion recognition using deep learning approach from audio–visual emotional big data," *Information Fusion*, vol. 49, pp. 69–78, 2019.

[54] V. Echeverria, R. Martinez-Maldonado, and S. Buckingham Shum, "Towards collaboration translucence: Giving meaning to multimodal group data," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, p. 39, ACM, 2019.

[55] M. N. Giannakos, K. Sharma, I. O. Pappas, V. Kostakos, and E. Velloso, "Multimodal data as a means to understand the learning experience," *International Journal of Information Management*, vol. 48, pp. 108–119, 2019.

[56] S. Kumar, M. Yadava, and P. P. Roy, "Fusion of eeg response and sentiment analysis of products review to predict customer satisfaction," *Information Fusion*, vol. 52, pp. 41–52, 2019.

[57] C. Park, P. Moghadam, S. Kim, A. Elfes, C. Fookes, and S. Sridharan, "Elastic lidar fusion: Dense map-centric continuous-time slam," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1206–1213, IEEE, 2018.

[58] C. Chou, A. Kingery, D. Wang, H. Li, and D. Song, "Encoder-camera-ground penetrating radar tri-sensor mapping for surface and subsurface transportation

infrastructure inspection," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1452–1457, IEEE, 2018.

[59] T. Bruls, W. Maddern, A. A. Morye, and P. Newman, "Mark yourself: Road marking segmentation via weakly-supervised annotations from multimodal data," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1863–1870, IEEE, 2018.

[60] S. Luo, W. Yuan, E. Adelson, A. G. Cohn, and R. Fuentes, "Vitac: Feature sharing between vision and tactile sensing for cloth texture recognition," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2722–2727, IEEE, 2018.

[61] T. Zhou and J. P. Wachs, "Early turn-taking prediction with spiking neural networks for human robot collaboration," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–9, IEEE, 2018.

[62] M. Di Castro, C. V. Almagro, G. Lunghi, R. Marin, M. Ferre, and A. Masi, "Tracking-based depth estimation of metallic pieces for robotic guidance," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5503–5508, IEEE, 2018.

[63] Z. Shao, H. Zhao, J. Xie, Y. Qu, Y. Guan, and J. Tan, "Unsupervised trajectory segmentation and promoting of multi-modal surgical demonstrations," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 777–782, IEEE, 2018.

[64] L. Santos, G. S. Martins, and J. Dias, "An extended bayesian user model (bum) for capturing cultural attributes with a social robot," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 320–325, IEEE, 2018.

[65] "Human-computer interaction: Why voice changes everything." `https://www.oculus.com/rift-s/features/`. (Accessed on 06/11/2019).

[66] A. Lim and H. G. Okuno, "The mei robot: towards using motherese to develop multimodal emotional intelligence," *IEEE Transactions on Autonomous Mental Development*, vol. 6, no. 2, pp. 126–138, 2014.

[67] T. Oka and K. Matsushima, "Multimodal manipulator control interface using speech and multi-touch gesture recognition," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, pp. 21–22, ACM, 2015.

[68] E. Bozkurt, Y. Yemez, and E. Erzin, "Multimodal analysis of speech and arm motion for prosody-driven synthesis of beat gestures," *Speech Communication*, vol. 85, pp. 29–42, 2016.

[69] J. Rico and S. Brewster, "Gesture and voice prototyping for early evaluations of social acceptability in multimodal interfaces," in *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, p. 16, ACM, 2010.

[70] N. Vidakis, A. Vlasopoulos, T. Kounalakis, P. Varchalamas, M. Dimitriou, G. Kalliatakis, E. Syntychakis, J. Christofakis, and G. Triantafyllidis, "Multimodal desktop interaction: The face-object-gesture-voice example," in *2013 18th International Conference on Digital Signal Processing (DSP)*, pp. 1–8, IEEE, 2013.

[71] O. Rogalla, M. Ehrenmann, R. Zollner, R. Becher, and R. Dillmann, "Using gesture and speech control for commanding a robot assistant," in *Proceedings. 11th IEEE International Workshop on Robot and Human Interactive Communication*, pp. 454–459, IEEE, 2002.

[72] S. Kang, K. Kim, and S. Chi, "Multimodal data fusion and intention recognition for horse riding simulators," in *Proceedings of the 2015 International Conference on Big Data Applications and Services*, pp. 296–297, ACM, 2015.

[73] D. Perzanowski, A. C. Schultz, W. Adams, E. Marsh, and M. Bugajska, "Building a multimodal human-robot interface," *IEEE intelligent systems*, vol. 16, no. 1, pp. 16–21, 2001.

[74] M. Skubic, D. Perzanowski, S. Blisard, A. Schultz, W. Adams, M. Bugajska, and D. Brock, "Spatial language for human-robot dialogs," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 34, no. 2, pp. 154–167, 2004.

[75] M. Turunen, J. Hakulinen, J. Hella, J.-P. Rajaniemi, A. Melto, E. Mäkinen, J. Rantala, T. Heimonen, T. Laivo, H. Soronen, *et al.*, "Multimodal media center interface based on speech, gestures and haptic feedback," in *IFIP Conference on Human-Computer Interaction*, pp. 54–57, Springer, 2009.

[76] T. Buffomante, "Rethink robotics." (2017).

[77] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE transactions on multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[78] J. F. Gorostiza, R. Barber, A. M. Khamis, M. Malfaz, R. Pacheco, R. Rivas, A. Corrales, E. Delgado, and M. A. Salichs, "Multimodal human-robot interaction framework for a personal robot," in *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*, pp. 39–44, IEEE, 2006.

[79] Z. Su, O. Kroemer, G. E. Loeb, G. S. Sukhatme, and S. Schaal, "Learning manipulation graphs from demonstrations using multimodal sensory signals,"

in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2758–2765, IEEE, 2018.

[80] J. Beil, I. Ehrenberger, C. Scherer, C. Mandery, and T. Asfour, "Human motion classification based on multi-modal sensor data for lower limb exoskeletons," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5431–5436, IEEE, 2018.

[81] D. Krupke, F. Steinicke, P. Lubos, Y. Jonetzko, M. Görner, and J. Zhang, "Comparison of multimodal heading and pointing gestures for co-located mixed reality human-robot interaction," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1–9, IEEE, 2018.

[82] "Who invented the computer?." `https://science.howstuffworks.com/innovation/inventions/who-invented-the-computer.htm`. (Accessed on 06/11/2019).

[83] "History of mobile cell phones." `https://bebusinessed.com/history/history-cell-phones/`. (Accessed on 06/12/2019).

[84] "Apple smart watch." `https://www.apple.com/apple-watch-series-4/`. (Accessed on 06/07/2019).

[85] "Microsoft modern keyboard with fingerprint id review." `https://www.techradar.com/reviews/microsoft-modern-keyboard-with-fingerprint-id`. (Accessed on 06/09/2019).

[86] "Sit back, relax, and enjoy a ride through the history of self-driving cars." `https://www.digitaltrends.com/cars/history-of-self-driving-cars-milestones/`. (Accessed on 06/12/2019).

[87] "These are the 11 coolest features of the tesla model s." `https://www.cnbc.com/2017/11/17/tesla-model-s-best-features.html`. (Accessed on 06/08/2019).

[88] "The pal-v liberty claims to be the first commercially available flying car." `https://www.theverge.com/2018/3/7/17091136/flying-car-pal-v-liberty-geneva-motor-show-2018`. (Accessed on 06/08/2019).

[89] L. M. Reeves, J. Lai, J. A. Larson, S. Oviatt, T. Balaji, S. Buisine, P. Collings, P. Cohen, B. Kraal, J. C. Martin, *et al.*, "Guidelines for multimodal user interface design," *Communications of the ACM*, vol. 47, no. 1, pp. 57–59, 2004.

[90] Senic, "The future of human computer interaction hardware startup knowledge medium," Jan 2015.

[91] Y. Xu, T. Price, J.-M. Frahm, and F. Monrose, "Virtual u: Defeating face liveness detection by building virtual models from your public photos," in *25th {USENIX} Security Symposium ({USENIX} Security 16)*, pp. 497–512, 2016.

[92] M. Urban and P. Bajcsy, "Fusion of voice, gesture, and human-computer interface controls for remotely operated robot," in *2005 7th International Conference on Information Fusion*, vol. 2, pp. 8–pp, IEEE, 2005.

[93] J. Preece, Y. Rogers, H. Sharp, D. Benyon, S. Holland, and T. Carey, *Human-computer interaction*. Addison-Wesley Longman Ltd., 1994.

[94] A. Tanaka and R. B. Knapp, "Multimodal interaction in music using the electromyogram and relative position sensing," in *Proceedings of the 2002 conference on New interfaces for musical expression*, pp. 1–6, National University of Singapore, 2002.

[95] K. Schwab, *The fourth industrial revolution*. Currency, 2017.

[96] O. Salman, M. F. A. Rasid, M. I. Saripan, and S. K. Subramaniam, "Multi-sources data fusion framework for remote triage prioritization in telehealth," *Journal of medical systems*, vol. 38, no. 9, p. 103, 2014.

[97] D. Gabriel, *Gesture Recognition using Electromyographic, Spatial and Temporal Input Data*. PhD thesis, 2016.

[98] G.-C. Luh, Y.-H. Ma, C.-J. Yen, and H.-A. Lin, "Muscle-gesture robot hand control based on semg signals with wavelet transform features and neural network classifier," in *2016 International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 2, pp. 627–632, IEEE, 2016.

[99] A. B. Csapo, H. Nagy, Á. Kristjánsson, and G. Wersényi, "Evaluation of human-myo gesture control capabilities in continuous search and select operations," in *2016 7th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, pp. 000415–000420, IEEE, 2016.

[100] A. Roitberg, N. Somani, A. Perzylo, M. Rickert, and A. Knoll, "Multimodal human activity recognition for industrial manufacturing processes in robotic workcells," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 259–266, ACM, 2015.

[101] S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz, "Fusion of face and speech data for person identity verification," *IEEE transactions on neural networks*, vol. 10, no. 5, pp. 1065–1074, 1999.

[102] E. Alpaydin, *Introduction to machine learning*. MIT press, 2009.

[103] G. King and L. Zeng, "Logistic regression in rare events data," *Political analysis*, vol. 9, no. 2, pp. 137–163, 2001.

[104] P. Xanthopoulos, P. M. Pardalos, and T. B. Trafalis, "Linear discriminant analysis," in *Robust data mining*, pp. 27–33, Springer, 2013.

[105] T. K. Ho, "Nearest neighbors in random subspaces," in *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pp. 640–648, Springer, 1998.

[106] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE transactions on systems, man, and cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.

[107] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pp. 338–345, Morgan Kaufmann Publishers Inc., 1995.

[108] C. Cortes and V. Vapnik, "Support vector machine," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[109] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *ICML*, 2014.

[110] G. Tesauro, "Temporal difference learning and td-gammon," *Communications of the ACM*, vol. 38, no. 3, pp. 58–68, 1995.

[111] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[112] H. H. Hoos and T. Stützle, "Towards a characterisation of the behaviour of stochastic local search algorithms for sat," *Artificial Intelligence*, vol. 112, no. 1-2, pp. 213–232, 1999.

[113] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[114] G. E. Uhlenbeck and L. S. Ornstein, "On the theory of the brownian motion," *Physical review*, vol. 36, no. 5, p. 823, 1930.

[115] R. P. Judd and A. B. Knasinski, "A technique to calibrate industrial robots with experimental verification," *IEEE Transactions on robotics and automation*, vol. 6, no. 1, pp. 20–30, 1990.

Appendix A

Application Code Snippets

In this appendix, We are going to present a few of the code fragments that contribute majorly to our application and are important in the development of connecting the Myo Arm band with a robotic arm.

Listing A.1: Myo Arm Band API

```
myo::Hub hub("com.jakechapeskie.SerialCommunication");

std::cout << "Attempting to find a Myo..." << std::endl;
myo::Myo* myo = hub.waitForMyo(10000);

if (!myo) {
throw std::runtime_error("Unable to find a Myo!");
}

std::cout << "Connected to a Myo armband!" << std::endl
<< std::endl;
DataCollector collector;
hub.addListener(&collector);
```

```
hub.setLockingPolicy(myo::Hub::LockingPolicy::
lockingPolicyStandard);
while (1) {
hub.run(1000 / 20);
collector.print();



if (strcmp(recvbuf, "Move Right") == 0)
{
printf("We are here %s\n", recvbuf);

if (SerialPortToDevice->IsOpen) {
        printf("We are here %s\n", recvbuf);
        for (int i = 0; i < 3; i++){
                SerialPortToDevice->WriteLine("waveOut");
                Sleep(300);
        }
}
ZeroMemory(recvbuf, recvbuflen);
//Sleep(1000);

}else if (strcmp(recvbuf, "Move Left") == 0)
{
if (SerialPortToDevice->IsOpen) {
        for (int i = 0; i < 3; i++){
                SerialPortToDevice->WriteLine("waveIn");
```

```
                    Sleep(300);
            }
    }
    ZeroMemory(recvbuf, recvbuflen);

    }else if (strcmp(recvbuf, "Move Up") == 0)
    {
    if (SerialPortToDevice->IsOpen) {
            for (int i = 0; i < 3; i++){
                    SerialPortToDevice->WriteLine
                    ("fingersSpread");
                    Sleep(300);
            }
    }
    ZeroMemory(recvbuf, recvbuflen);

    }else if (strcmp(recvbuf, "Move Down") == 0)
    {
    if (SerialPortToDevice->IsOpen) {
            for (int i = 0; i < 3; i++){
                    SerialPortToDevice->WriteLine("fist");
                    Sleep(300);
            }
    }
    ZeroMemory(recvbuf, recvbuflen);
```

```
}else if (strcmp(recvbuf, "Move Gripper") == 0)
{
if (SerialPortToDevice->IsOpen) {
        for (int i = 0; i < 8; i++) {
                SerialPortToDevice->WriteLine
                ("doubleTap");
                Sleep(300);
        }

}
ZeroMemory(recvbuf, recvbuflen);

}
else {
std::string poseString = (collector.currentPose.
toString());
String^ poseStrorageString = gcnew String(
poseString.c_str());
printf("Only gesture: %s\n", poseStrorageString);
if (SerialPortToDevice->IsOpen) {
        SerialPortToDevice->WriteLine
        (poseStrorageString);
        poseStrorageString = "";
}
}
```