PROJECT SANWADA

INTELLIGENT MOBILE ASSISTANT FOR HEARING IMPAIRERS TO INTERACT WITH THE SOCIETY

Project ID: 17-092

Project Final Report

S. Y. M. Perera (IT14029264)

B.Sc. Special (Hons) in Information Technology

Department of Information Technology

Sri Lanka Institute of Information Technology

Sri Lanka

Submitted on 04/10/2017

SANWADHA

INTELLIGENT ASSISTANT FOR HEARING IMPAIRERS TO INTERACT WITH THE SOCIETY

S. Y. M. Perera

IT14029264

B.Sc. Special (Hons) in Information Technology

Department of Information Technology

Sri Lanka Institute of Information Technology

Sri Lanka

4th October 2017

DECLARATION

I declare that this is my own work and this final document does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of our knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Author:

Student ID	Name	Signature
IT14029264	S. Y. M. Perera	

ACKNOWLEDGMENT

I would like to express our sincere gratitude to our advisor Prof. Samantha Thelijjagoda for the continuous support of our research, for his patience motivation, enthusiasm, and immense knowledge. His guidance helped us in all time of research and writing of this thesis. Besides our advisor, I would like to thank our thesis board: Mr. Jayantha Amararachchi for his encouragement, insightful comments and hard questions. An excessive thank goes to Mr. Kasun Jayathunga, Founder & CEO at "Ahanna

Foundation" who provided insight and expertise that greatly assisted the research in gathering facts about hearing-impaired community.

This research was in part supported by Ms. Rajitha Viduransi, Interpreter for the Deaf, "Ahanna Foundation" who was the foremost cause for turning the research problem towards deaf community and supports to get the idea about sign language used by hearing-impaired people.

Then again special thank goes to Rev. Sister Nirmali, Principal - St. Joseph's School for the Deaf, Ragama who allowed us to realize the real situation about the problems faced by hearing-impaired people.

ABSTRACT

As we all are Sri Lankans we are the only people who use the Sinhala language over the world. Consequently, there is a great demand for a TTS system in Sinhala language particularly a real-time application for visually impaired people. Although regular users or document readers will get the benefit of a document reader in there tight day today schedules. The Internet has taken communication to unprecedented heights today. People all around the world use the Internet to get connected and communicate with those from around the world. These include methods such as emailing, community sites and most importantly chat systems. Using chat systems to communicate has become a trend and the most fashionable way to connect with people all around the world. However, these privileges of the Internet are limited only to people who are normal and are abled. But people who are hearing-impaired are isolated and denied these uses of the Internet. To help the hearing-impaired with their communication the "Sanwadha" chat system has done come up. This chat application will include different means of communication other than the conventional text to text keyboard conversation. Unlike an ordinary chat system, the "Sanwadha" chat system is intelligent to determine the mode of communication of the user. The user in this context means a hearing-impaired person, according to disability. The "Sanwadha" chat system is a collection of technologies that already exist, technologies such as the conversion of voice to text and vice versa and text to text, and new incorporates technologies such as the conversion of sign language. This document describes the "Sanwadha" System and its impact on society today. It indicates clearly the background of the "Sanwadha" system using literature survey and the problem which must overcome by the "Sanwadha" team.

Keywords: *Hearing-impairers, Instant Messaging, Mobile Application, Voice recognition, Natural Language processing, Graphic Interchange Format Introduction*

Table of Contents

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Background

In 2003, a research has been carried out about the number of deaf people in Brussels. The results of this study showed that about 1 on 1000 people are deaf. It wondered how Hearing-impairers deal with their situation daily and if there are difficulties with the hearing society. Different sources are referred by us and it was immediately clear that the life of a Hearing-impairer is not that much easy. Research has shown several discoveries concerning Deaf community. Most of the time Hearing-impairers discriminated and excluded by the society. It has been recognized that the hearing society has misperceptions about Hearing-impairers [1].

Problems can be found in several environments: at work, at school, in medical world, in social life etc. Due to the communication problems, Hearing-impairers face many barriers. Deafness is invisible disability which is not surprising. Because you can't see if a person is Deaf [2], [40].

Let's consider about the evolution in communication, in the last half of the 20th century the communication developed rapidly. Today this development has made communication in the day-to-day life easy. With the advancement of the technology, accessing internet has become the most imperative thing. Thus, people are more prone to use the internet as a means of communication more frequently. Using chat systems for communication has become a trend and the most popular way to connect with people all around the world. Although this is the case, today this facility is restricted only to ordinary people. Yet people who are differently-abled are isolated and denied of this facility just because of their disability. Per our knowledge and experience, most of the chat systems are based on text based chatting. Have you ever thought about, how a person with a disability uses this kind of application? Using new technologies to help people with disabilities is highly regarded and much research in this area is underway [3], [38].

The focus of this investigation goes towards the Deaf community. The technology has not sufficiently reached to the Hearing-impairs. If they want to use these kind of chat applications, those applications should support the ways that deaf people can manage. Although visually-impaired people can communicate by using the human language. Hearing-impairers cannot use that language. They usually comfortable with the sign language. Hence, these applications should support the sign language [3], [39].

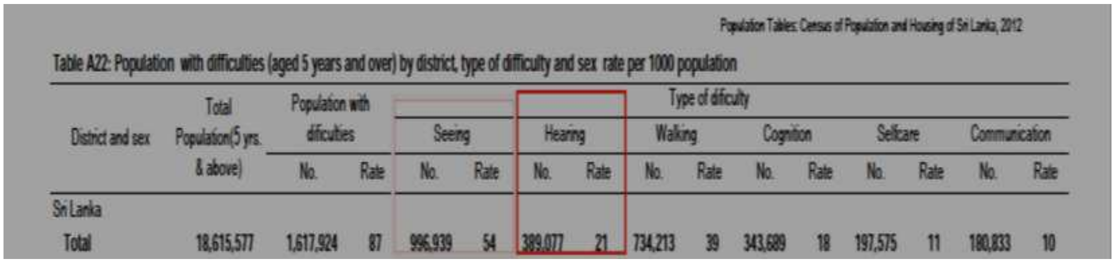
Sign languages are natural languages with their own grammar and syntax, specially formulated for the deaf people. Make use of finger spellings, body language, lip pattern and manual communication, to convey the meaning. It mainly involves the use of orientation and movement of hands. The language can be taught only by a person who is specially trained in it. Today, the 'differently-able' people can communicate to the rest of the world as easily and effectively as the able bodied. The credit goes to the sign language which was developed earlier. Most countries have their own national sign languages. Sri Lankan Deaf community also using Sinhala sign language [4], [37].

Nowadays few different applications for people using English sign languages and other sign languages. Our main aim is to reach the Sri Lankan Deaf community who are using Sinhala sign language. Besides Sinhala is the foremost language in Sri Lanka. Today communication of Hearing-impairers with ordinary people are done by an Interpreter. Furthermore Hearing-impairers comfortable with Lip reading. Yet with the absence of an Interpreter, there is a huge gap between Hearing-impairers and Hearing people.

In the investigation towards Deaf community, Ragama School for Deaf was the most imperative place visited. Along with that the Interpreter from “Ahanna” community supports us to reach our background study very effectually. The conclusion gained was that it is equally important that hearing people should learn to deal with Hearing-impairers.

1.2. Literature Survey

Agreeing to the Census of Population and Housing – 2012 there are 389,077 hearing impaired people out of 18,615,577 people who are above 5 years old. That is roughly 21 people out of 1000 population are hearing impaired in Sri Lanka [4]. The deafness is variable. It can occur at any stage of life cycle, it may impact on the individual’s ability to function on a day-to-day basis and it may or may not be disabling.

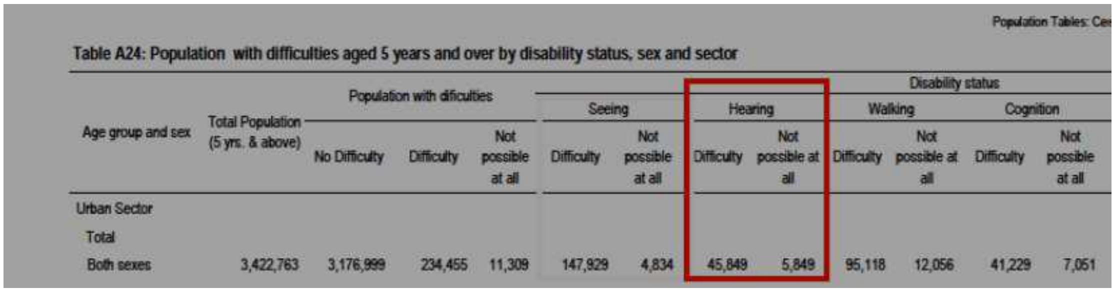


Population Tables: Census of Population and Housing of Sri Lanka, 2012

Table A22: Population with difficulties (aged 5 years and over) by district, type of difficulty and sex rate per 1000 population

District and sex	Total Population(5 yrs. & above)	Population with dificulties		Type of dificulty											
				Seeing		Hearing		Walking		Cognition		Selfcare		Communication	
		No.	Rate	No.	Rate	No.	Rate	No.	Rate	No.	Rate	No.	Rate	No.	Rate
Sri Lanka															
Total	18,615,577	1,617,924	87	996,939	54	389,077	21	734,213	39	343,689	18	197,575	11	180,833	10

Figure 1.1: Population with difficulties rate per 1000 population



Population Tables: Ce

Table A24: Population with difficulties aged 5 years and over by disability status, sex and sector

Age group and sex	Total Population (5 yrs. & above)	Population with dificulties			Disability status							
					Seeing		Hearing		Walking		Cognition	
		No Difficulty	Difficulty	Not possible at all	Difficulty	Not possible at all	Difficulty	Not possible at all	Difficulty	Not possible at all	Difficulty	Not possible at all
Urban Sector												
Total												
Both sexes	3,422,763	3,176,999	234,455	11,309	147,929	4,834	45,849	5,849	95,118	12,056	41,229	7,051

Figure 1.2: Population with difficulties

This research is mainly focusing on these people. Hearing and hearing impaired people are having difficulties when they are communicating in day to day life. These people are using sign languages to communicate with each other." A sign language is a language which chiefly uses manual communication and body language to convey meaning, as opposed to acoustically conveyed sound patterns. This can involve simultaneously combining hand shapes, orientation and movement of the hands, arms or body, and facial expressions to fluidly express a speaker's thoughts" [Y]4p. There are around three hundred sign languages in the world, each sign language differs from each other by country and the language. Hand gestures in sign languages are defined for both alphabets and word phrases.

Some of the sign languages used around the world are:

- Sri Lankan Sign Language
- American Sign Language
- British Sign Language
- Italian Sign language
- French sign languages

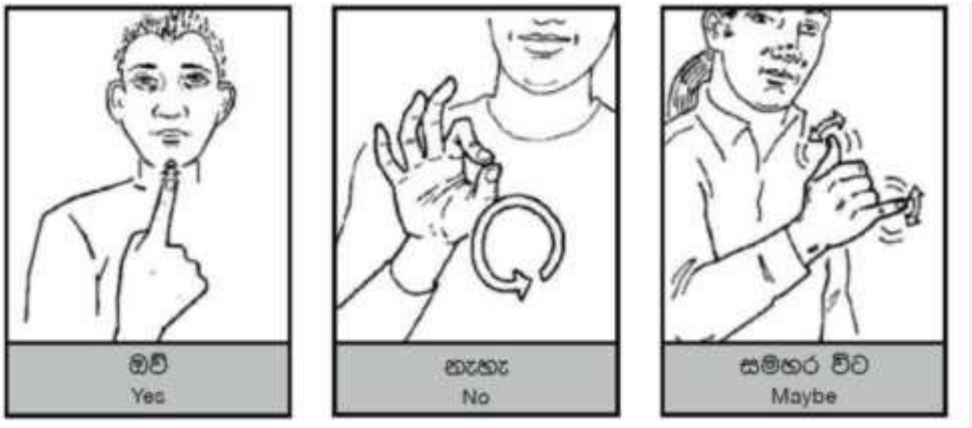


Figure 1.3: Sri Lankan sign language

There are few different applications for people using English sign languages and other sign languages. Research is mainly focusing on hearing and speech impaired people in Sri Lanka who are using Sinhala sign language and when they are communicating with Sinhala. One difficulty hearing and hearing impaired people are facing is they can only guess what a speaker is saying by lip reading, to understand totally they need someone to interpret in sign language what speaker is saying. Another difficulty is when trying to communicate something to a person who doesn't know sign language. They need interpreter to translate sign language to native speaking language. Conversational speech can be measured as having a Loudness of approximately 60 decibels (dB). Hearing is considered significantly restricted when the ear cannot interpret or process sounds of 25 dB or more [5].

The most natural way to communicate for human beings is through words. When considering about the experience of disability most of them don't have experience of expressing their thoughts with hearing people. High percentage of disabled people are Hearing-impairers. [1].

There are 75 million of deaf people use sign language as their first language. Each country has one or sometimes two or more sign languages. There are some common techniques used by deaf people to communicate with normal people. Some deaf people use speech or sign language only or a combination, some may use finger spelling or writing or body language and facial expressions. Like spoken languages, signed languages vary. Sign languages have their own accents, dialects, and idiosyncratic vocabulary. Signs may be limited to regions, schools, or even families [6].

As a help to these people so far, many applications, systems and devices have been introduced. But the main problem is to connect both non-hearing impaired with hearing impaired simultaneously. For that there are only limited number of systems are evolved. There are applications which can only turn voice to text or sign language to text separately. This research is introduced this system as a two-way communication system. These kinds of systems mainly focused on accuracy level. When working with hand gesture recognition part, many systems used image processing as the technique. It takes lots of time and lots of processing to work on. Even though it achieved, when it comes with hand gestures some rotational movements cannot be track with image processing [7].

The Bolt, Beranek and Newman System

The first computer-based speech training aid was developed around a Digital Equipment Corporation PDP-8E minicomputer. This was an experimental system, resulted directly from its development. The system consisted of 3 sensors (voice-microphone, accelerometer on the throat, and accelerometer on the nose) a preprocessor, the computer, and various output displays. The preprocessor included a pitch extractor, a spectrum analyzer, and a nasal detector [8].

Hand Gesture Recognition

This describes using RGB color spaces and models and presents, some possible ways of segmentation with algorithms. Various experiments were conducted for different gestures and results were obtained with accuracy. The algorithms were implemented in MATLAB programming language. Here it concluded Capturing the hand without the glove results in inaccurate outputs. Data base creation and testing using a GUI makes the system more user friendly. The database can be expanded with more number of hand gestures and its different possibilities to improve the performance of the system. This system consists of a basic web camera which points to the signer, MATLAB -which performs the image processing operations and an audio speaker or a display to convey the message shown by the signer. Here a colored glove is used by signer. The gloves will have red, blue, green color pattern on each finger. The intensity of the color changes with gestures. The gestures are captured by a camera. The intensity changes of the colors are detected. The gestures are detected with image processing using MATLAB [9].

Speech to Text Conversion in Real-time

This software is developed to enhance user's way of speech through correctness of pronunciation following the English phonetics. This desktop software allows one to learn, judge and recognize their pronunciation in English language. This also provide an extra add-on feature which enhance the user's communication skills by an option of text to speech conversion also. This software presents method to design a Text to Speech conversion module using Mat lab and visual studio. As a real time system, this provides a good timing (within 2-3 seconds) and less cost when compare to other voice to text converting systems. Yet this system only can be used with American accent and this is a desktop application [10].

Analysis and selection of features for gesture recognition based on a micro wearable device

This one is considering the flexibility of human finger, a device is developed which can be put it on a finger to detect the finger gestures, and 12 kinds of one-stroke finger gestures are defined per the sensing characteristic of the accelerometer. Designed a wearable device with an accelerometer to wear on finger and catch movements in 3D space. Experiment results indicate the feature subset can get satisfactory classification results of 90.08% accuracy using 12 features considering the recognition accuracy and dimension of feature set. The system is a ring shape sensing device based on a 3-trial accelerometer. To the system adopt the algorithm of feature selection, stepwise regression. This system defines great accuracy level even with the gesture combinations. But with this system only can be performed very simple set of gestures plus this is not a portable system [11].

Recognition of no manual markers in American Sign Language (ASL) using non-parametric adaptive 2D-3D face tracking

This one address the problem of automatically recognizing linguistically significant non-manual expressions in American Sign Language from video. Develop a fully automatic system that can track facial expressions and head movements, detect and recognize facial events continuously from video. The main contributions of the proposed framework are the following:

- Built a stochastic and adaptive ensemble of face trackers to address factors resulting in lost face track.

- Combine 2D and 3D deformable face models to warp input frames, thus correcting for any variation in facial appearance resulting from changes in 3D head pose.
- Use a combination of geometric features and texture features extracted from a canonical frontal representation. The proposed new framework makes it possible to detect grammatically significant non-manual expressions from continuous signing and to differentiate successfully among linguistically significant expressions that involve subtle differences in appearance [12].

Most of these systems and devices are only focused on a one side communication. But in this application, both focused on text to sign language and sign language to text. The specialty of this system is there is no such a system invented and not for Sinhala language.

Observation from the Literature Review,

- There are very less number of systems have been introduced for Sinhala language.
- The systems which are using image processing techniques are hard to implement and cannot reach the higher accuracy levels of detection when it comes to rotational of gestures.
- Some systems are high cost and technology level is not tally with our country.
- Lacks the expertise and the capacity to deal with deaf people to train them.
- Not having enough existing systems to use for the deaf users with well based manner and remains drawbacks of them.

1.3. Research Gap

There is a communication gap between Hearing-impairers and the ordinary people. Most of the time that is being filled through an Interpreter. It would be a problem when there is no Interpreter. By now, there are some solutions to cover this problem. But those solutions couldn’t reach the Sri Lankan Deaf community. Most of them are not flexible with the Deaf users and they are not casing all the extents they need. So, by today Hearing-impairers have challenged with a huge communication gap in their day to day life. Our proposed application would be the finest solution for this gap.

1.3.1. Research gap in Creating chat application

The purpose of this project is to design and implement a multi featured chat application among ordinary people and hearing impairers. This chat application would be included by different means of communication other than the conventional text to text keyboard conversation. Things such as interpreting sign language signs to text and voice to text will be the most useful features of the system. Ordinary user could be more comfortable with Sinhala and Singlish texting. Moreover, this application can use with mobile data and if there are no mobile data, such a case user can use offline message feature.

1.3.2. Research gap in Voice Recognition

There has been a significant amount of research done in voice recognition where a system can be trained to identify a variety of accents based on various voice models which are trained to identify voice. A drawback found is the problem to recognize voice in a noisy environment. This of course cannot be eliminated 100% and specialized have been created to avoid the above-mentioned problem. But a perfect software solution has not been invented yet. Another problem unique to Sri Lankan users is because Sri Lankans have a unique accent when speaking English. Voice recognition systems fail to detect some words pronounced by Sri Lankans. Along with that there is a need to address this problem with our research. Voice recognition systems also take high processing power and the motive to reduce processing power is another area of concern.

1.3.3. Research gap in creating 2D Hand Model

This propose a real-time model-based 2D hand tracker that combines image regions 2-axis accelerometer placed on the user's hand. The accelerometer and tracker are synchronized by casting the calibration problem as one of principal component analysis. Based on the assumption that often, the number of possible hand configurations is limited by the activity the hand is engaging in. Use a multiclass pose classifier to distinguish between a few activities dependent articulated hand configurations [9].

1.3.4. Research gap in Semantic Analysis

Identifying Semantic analysis would get the meaning of a set of words and convert that meaning into a GIF. Enable the user get the core idea of the message without having nonsense words [17].

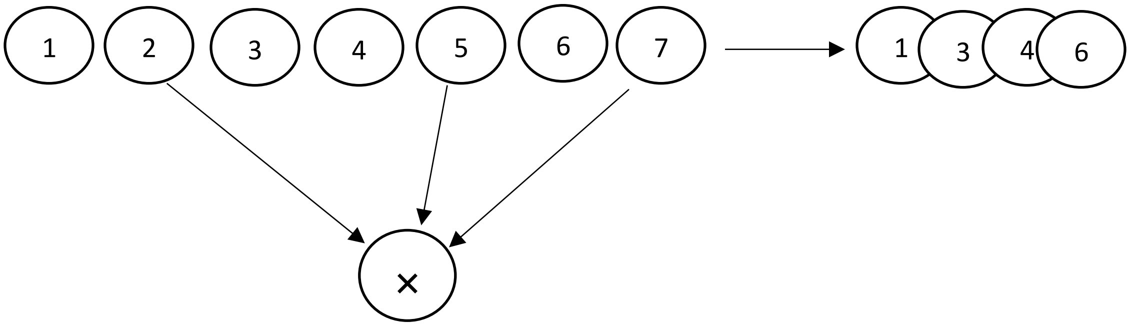


Figure 1.1: Semantic Analysis procedure

1.3.5. Drawbacks of current solutions available

Drawbacks of the solutions available today can be summarized as follows

Evaluation Study in Diverse

Deaf chat

Deaf Chat facilitates communication between Deaf and Hearing individuals. It replaces the pencil and paper that is frequently used, plus you can communicate over moderate distances.

A network connection is established between two devices (phones or tablets). The first individual can input text via voice recognition or the keyboard into his Local text area and send this to the second device. On the second device, the text will appear in the Remote text area. The second individual can respond back to the first by entering text into his Local text area, again using either voice recognition or the keyboard.

Using a network connection rather than Bluetooth allows the individuals to be near each other or separated by a large distance. The network connection will most likely be Wi-Fi, but it could be an Intranet or even a connection via the Internet [13].

Deaf - Hearing chat

DH Chat is a system for face-to-face communication between deaf and hearing people without a sign interpreter. If you are a hearing person, you can communicate with your deaf relatives, friends, clients, employees and so on. If you are a deaf person, you can make a face-to-face conversation with hearing people without sign interpreter. You can use the system everywhere: at home, at your work, at restaurants, during your education and so on [14].

NGTS

NGTS (Next Generation Text Service) is a fantastic application for helping deaf and hard of hearing people to communicate over the phone via a text relay assistant. NGTS is especially handy for using at work and can be tailored to meet your specific communication needs. User can choose from type and read, speak and read, type and hear, speak and hear options, and it's really simple to use.

Glide

Glide – Video Chat Messenger is a deaf person's favorite. The famous video messaging app allows you to send super-fast videos up to 5 minutes long, and completely hands-free. Other elements include group chats and uploading videos to social media.

BizzBook

BizzBook application lets user connect with local business using Text Messaging. User can talk to businesses. For example if user need to know their specials or hours of operation- or Group Chat. z5 Mobile is a video relay service. Whenever user need to speak with a non-signer this is the way to go. Z5 Mobile worked over Video with translators who verbally communicate your messages to the hearing caller.

Evaluation Study in Sri Lanka

Nihanda System

This system used for children who are diagnoses with hearing impaired. Used leap motion controller to track signs and convert them to voice. They implemented game based learning system to hearing impaired children to learn sign language easily. System demonstrate how to identify individual signs and phonetics though videos and images. Mind teaser games uses to self-motivate children to improve their learning abilities. This system capture voice and gives 2D images.

Ahanna System

"Ahanna" is mainly focused on teaching the Sinhala sign language to the users who uses that system. It is web based online application.

The main intention of the "Ahanna" is to spread the pure Buddhism to the deaf community of Sri Lanka through Sri Lankan Sign Language (SLSL) while gifting many more valuable activities, innovative products & new ideas to improve the knowledge, education & quality of Sri Lankan Deaf Community [15].

KATHANA Sinhala speech recognition system

"KATHANA" is a solution for recognizing and interpreting voice. Application converts an acoustic signal which represents human speech done in Sinhala language captured by a microphone, to a set of words. Emphasis is that this acoustic wave represents a human speech done in Sinhala language. The recognized words which are the results can be used for applications as commands, data entries or could be served as the input to further linguistic processing to achieve speech understanding [16].

Features	**Deaf chat**	**Deaf hearing chat**	**Nihanda**	**Ahanna**	**Kathana**	**Sanwada**
Speech to sign translation-Sinhala	✗	✗	✗	✗	✗	✓
Text to Sign language – Sinhala & Singlish	✗	✗	✗	✗	✓	✓
Translated sign language to GIF	✗	✗	✗	✗	✗	✓
Sign language using 2D modeling	✗	✓	✗	✗	✗	✓
Price/Open source	$ 0.99	$ 2.99	FREE	FREE	FREE	--
Stickers and animated stickers	✓	✗	✗	✗	✗	✓
Interaction with Facebook messenger	✗	✗	✗	✗	✗	✓
Mobile application	✓	✓	✗	✗	✗	✓

Table 1.1: Comparison with Available system

1.4. Research Problem

- Communication between each other is one of the most essential thing to every human being but unfortunately hearing impaired people are having difficulties in communicating with day to day life in the society.
- Subsequently it is essential to bridge the communication gap between hearing-impaired people and the ordinary people.
- Deaf people communicate visually and physically rather than audibly. Many deaf people feel awkward or become frustrated trying to communicate with ordinary people, especially when no interpreter is available.
- When consider about deaf people in distance; there's no way to share emotions and feelings unless they meet each other.
- Deaf community discourage to be social. They do not have any desire to meet each other and share ideas.
- Due to having communication problem; there are many concerns faced by deaf people in day to day travel once that person does not know how to go.
- When following the day to day scenarios; deaf people unable to get any support from ordinary people since there isn't any common communication mode.

This is the problem addressed in our research. Several researches have tried to address this issue, although none of them could not grasp the achievement successfully.

1.5. Objectives

1.5.1. Main Objectives

- The main intention of the investigation is to deliver excessive support by enabling hearing impaired people to communicate with others, share feelings and ideas, actively interact with the society and help that they require with minimum amount of effort and time. And, allowing the hearing impairs to play the role by way of ordinary people without having desertions.

- To influence the Deaf community with the highest technology evolution to make hearing-impairers more comfortable in the global world.

1.5.1. Specific Objectives

- To identify many Hearing-impairers prefer to communicate with Sign languages which leads to the communication barrier with ordinary people. Subsequently Deaf people become frustrated to interact with the society.
- To establish the pathway for investigation, research papers and documents were surveyed mostly. Furthermore, the real obligation was discovered through the interviews conducted with students of Deaf School.
- To determine the use of mobile applications for deaf people can be observed as a diligence that allows them regardless to utilize to any need of learning and communication at any time anywhere.
- To emerge the application in Sinhala language to reach the Sri Lankan deaf community in an effective way.
- To advance the text message to a Graphic Interchange Format (GIF) to get the message in sign language with more accurate and attractive manner.

- To allow the generation of own sign language using 2D model provided which makes hearing impairers more comprehend about the message they want to direct.
- To enhance Sinhala voice recognition algorithm.
- To interact with the most popular social media like Facebook Messenger.
- To verify that the product is reliable for Hearing-impaired community to lead to a sociable life.

Apart from above mentioned objectives following intentions could be identified.

- To minimize the Barriers in communication

The main problem that the differently able people face is the communication difficulty. These barriers affect access to public information, opportunities to express oneself and access to essential services such as health, housing, transportation, education and employment. Even though we can't address all these areas, our system is an effort to minimize the barriers between differently able people and normal people in communication, by providing them a way to interact with other people and the society.

- To eliminate Barriers in education

People with disabilities often have access to less and inferior education than people without disabilities, because of many types of barriers. This system will provide them a good opportunity to share their knowledge and experiences with other each other.

- To minimize Barriers in healthcare

There are numerous barriers encountered in the access and delivery of health services. The major unavoidable issues in health care are physical access issues, funding, attitudinal and communication issues. Our system can address most of these areas other than funding and attitudinal as they are depending on the person. The users can connect with their doctors via system and can get the opinion.

- To minimize Barriers in developing human relationships

Isolation of differently able people in today's world has become a severe problem even though the modern and developed world today has failed to realize. Most of the time people who are differently able, not care for by their families, alone in places under the care of never met charity workers, are left out to be so alone. By using this system these people may able to interact with the outside world and develop good relationships with them.

- To be an outstanding team player

As a team of four dynamic individuals collaborating to achieve established goals and objectives of the CDAP project process by us. Inbuilt team-player potential to produce outstanding personage are hoped to recognize and enhance.

1.6. Research Questions

- What are the features that hearing-impaired person expects from a mobile application?
- What are the social media services that are connected?
- What are the languages that used for the input text?
- How to deliver the GIF message to the user?
- What are the technologies worked out?
- What are the techniques that can make the best performance?

2. RESEARCH METHODOLOGY

This chapter illustrates the methodology for handling the project. It’s a methodical approach to the research, gathering requirements, designing and implementation to create effective solution to an existing problem an area where improvement is required.

Proposed solution presents an intelligent assistant for hearing impairers to interact with the society.

The project has a very significant research areas like, Natural Language Processing (NLP), Voice Detection, Machine learning, Artificial Intelligence, Graphic Interchange Format (GIF) conversion and Mobile platform development. Machine Learning and GIF conversion is important for the identification of individual words in each Text and converted sign language send via compressed GIF files. Research conducted further study on above mentioned research areas then the information can be used to achieve the objectives [26].

2.1. System Overview

Considering the outcome of the literature review, it is conceivable to decide the most appropriate tools, technologies and software solutions for the implementation phase. In some cases of design conclusions, study more than one possible technologies and take performance and dependencies into deliberation.

The projected solution can be divided to following key components:

- 2D Model creation
- Text Conversion Mechanism
- GIF file Compression and Extraction Mechanism
- Voice Recognition Module

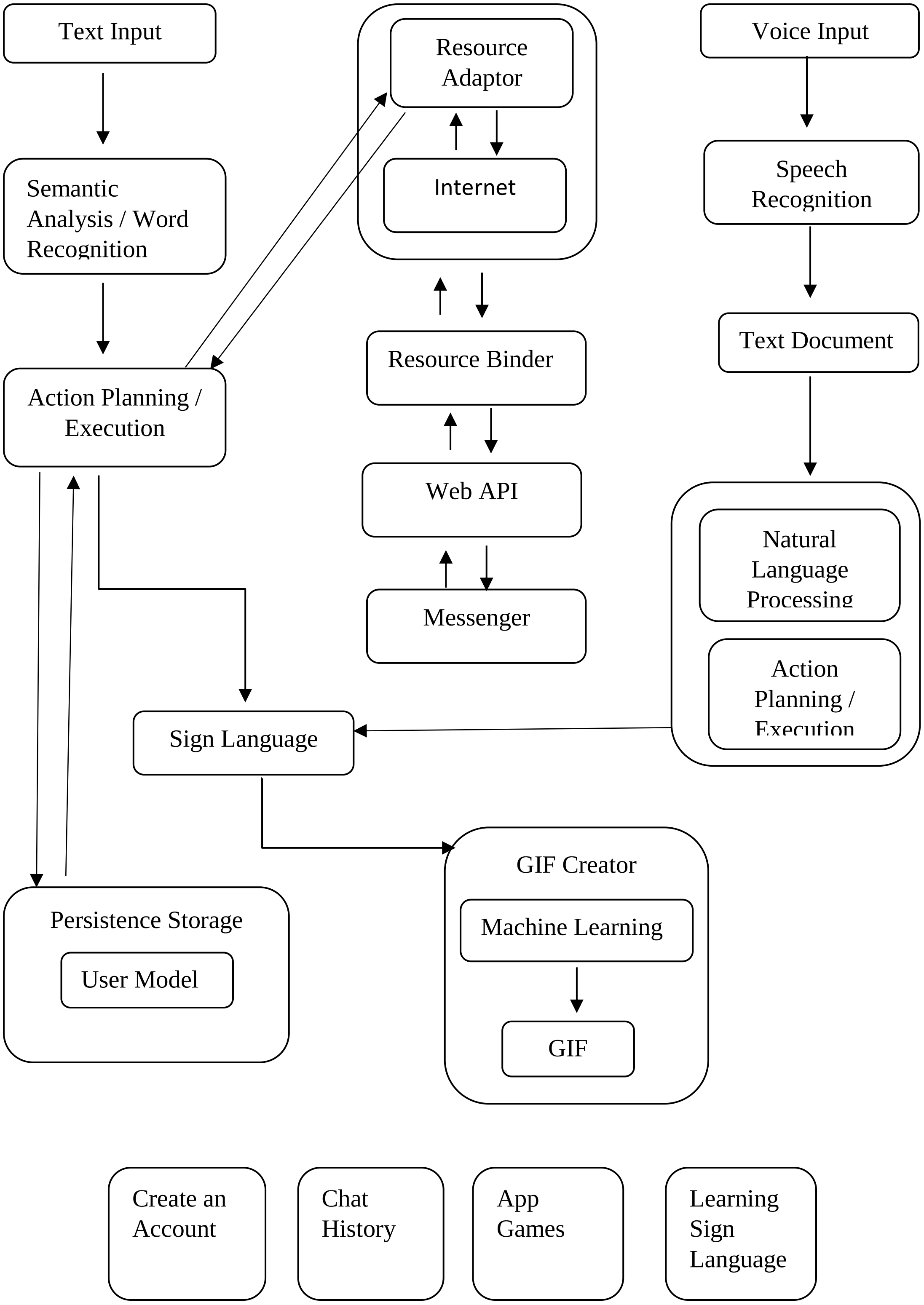


Figure 3.1: System overview of proposed solution

2.1.1 Voice Recognition

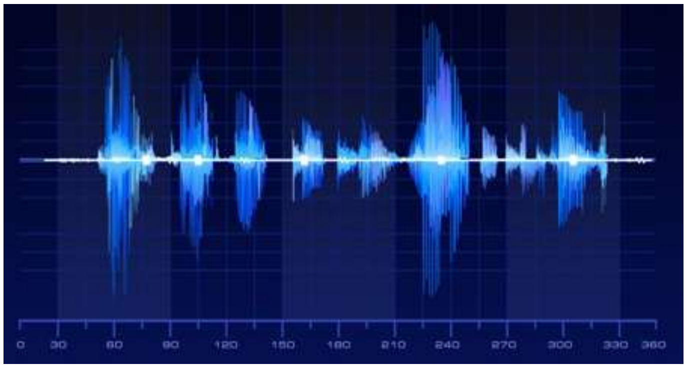


Figure 2.2: Voice Recognition

A speech recognition engine (or speech recognizer) takes an audio stream as input and turns it into a text transcription. The speech recognition process can be thought of as having a front end and a back end.

The front end processes the audio stream, isolating segments of sound that are probably speech and converting them into a series of numeric values that characterize the vocal sounds in the signal.

This human ability has inspired researchers to develop systems that would emulate such ability. From phoneticians to engineers, researchers have been working on several fronts to decode most of the information from the speech signal. Some of these fronts include tasks like identifying speakers by the voice, detecting the language being spoken, transcribing speech, translating speech, and understanding speech.

Among all speech tasks, (ASR) has been the focus of many researchers for several decades. In this task, the linguistic message is the information of interest. Speech recognition applications range from dictating a text to generating subtitles in real-time for a television broadcast. Despite the human ability, researchers learned that extracting information from speech is not a straightforward process. The variability in speech due to linguistic, physiologic, and environmental factors challenges researchers to reliably extract relevant information from the speech signal. In spite of all the challenges, researchers have made significant advances in the technology so that it is possible to develop speech-enabled applications.

Mathematical Formulation

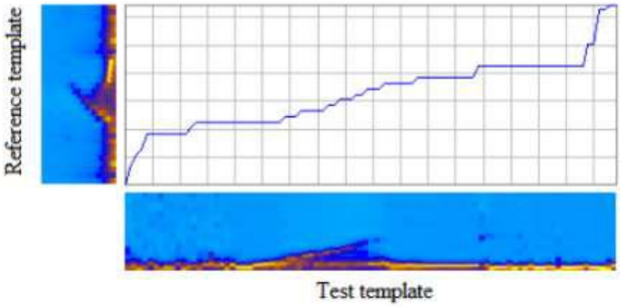
The speech recognition problem can be described as a function that defines a mapping from the acoustic evidence to a single or a sequence of words. Let X = (x_1, x_2, x_3, …, x_t) represent the acoustic evidence that is generated in time (indicated by the index t) from a given speech signal and belong to the complete set of acoustic sequences, χ. Let W = (w_1, w_2, w_3, …, w_n) denote a sequence of n words, each belonging to a fixed and known set of possible words, ω. There are two frameworks to describe the speech recognition function: template and statistic.

Template Framework

In the template framework, the recognition is performed by finding the possible sequence of words W that minimizes a distance function between the acoustic evidence X and a sequence of word reference patterns (templates) [1]. So the problem is to find the optimum sequence of template patterns, R * , that best matches X, as follows

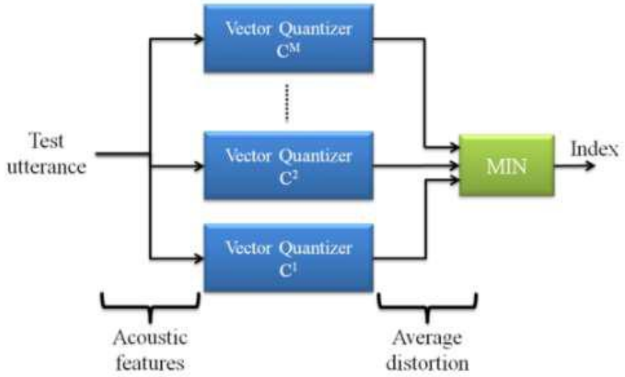
$$R^{*} = \underset{R^{s}}{\operatorname{argmin}}\, d(R^{s}, X)$$

where R^S is a concatenated sequence of template patterns from some admissible sequence of words. Note that the complexity of this approach grows exponentially with the length of the sequence of words W. In addition, the sequence of template patterns does not take into account the silence or the coarticulation between words. Restricting the number of words in a sequence [1], performing incremental processing [2], or adding a grammar (language model) [3] were some of the approaches used to reduce the complexity of the recognizer.



dynamic time warping of two renditions of the word "one"

The VQ method encodes the speech patterns from the set of possible words into a smaller set of vectors to perform pattern matching. The training data from each word $w_i \in \omega$ is partitioned into M clusters so that it minimizes some distortion measure [1]. The cluster centroids (codewords) are used to represent the word w_i , and the set of them is referred to as codebook. During recognition, the acoustic evidence of the test utterance is matched against every codebook using the same distortion measure. The test utterance is recognized as the word whose codebook match resulted in the smallest average distortion. Fig. 2 illustrates an example of VQ-based isolated word recognizer, where the index of the codebook with smallest average distortion defines the recognized word. Given the variability in the speech signal due to environmental, speaker, and channel effects, the size of the codebooks can become nontrivial for storage. Another problem is to select the distortion measure and the number of codewords that is sufficient to discriminate different speech patterns.



VQ-based isolated word recognizer

Statistical Framework

In the statistical framework, the recognizer selects the sequence of words that is more likely to be produced given the observed acoustic evidence. Let $P(W|X)$ denote the probability that the words W were spoken given that the acoustic evidence X was observed. The recognizer should select the sequence of words W satisfying

$$\widehat{W} = \underset{W\in\omega}{\operatorname{argmax}} P(W|X).$$

However, since $P\ W\ X$ is difficult to model directly, Bayes‘rule allows us to rewrite such probability as

$$P(W|X) = \frac{P(W)P(X|W)}{P(X)}$$

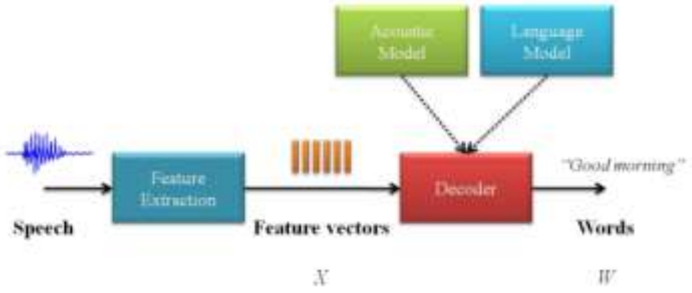
where P W is the probability that the sequence of words W will be uttered, P$(X|W)$ is the probability of observing the acoustic evidence X when the speaker utters W, and $P(X)$ is the probability that the acoustic evidence X will be observed. The term $P(X)$ can be dropped because it is a constant under the max operation. Then, the recognizer should select the sequence of words W that maximizes the product $P(W)$ $P(X|W)$, i.e.,

$$\widehat{W} = \underset{W \in \omega}{\operatorname{argmax}} P(W)P(X|W). \qquad (1)$$

This framework has dominated the development of speech recognition systems since the 1980s.

Speech Recognition Architecture

Most successful speech recognition systems are based on the statistical framework. Equation (1) establishes the components of a speech recognizer. The prior probability P(W) is determined by a language model, and the likelihood P $(X|W)$ is determined by a set of acoustic models, and the process of searching over all possible sequence of words W that maximizes the product is performed by the decoder.



Automatic Speech Recognition Classification

ASR systems can be classified according to some parameters that are related to the task. Some of the parameters are:

Vocabulary size: speech recognition is easier when the vocabulary to recognize is smaller. For example, the task of recognizing digits (10 words) is relatively easier when compared to tasks like transcribing broadcast news or telephone conversations that involve vocabularies of thousands of words. There are no established definitions, but small vocabulary is measure in tens of words, medium in hundreds of words, large in thousands of words and up [6]. However, the vocabulary size is not a reliable measure of task complexity [7]. The grammar constraints of the task can also affect the complexity of the system. That is, tasks with no grammar constraints are usually more complex because all words can follow any word.

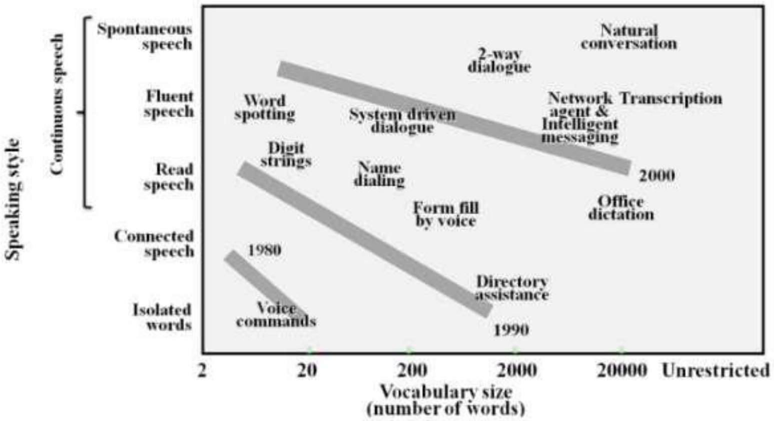
Speaking style: this defines whether the task is to recognize isolated words or continuous speech. In isolated word (e.g., digit recognition) or connected word (e.g., sequence of digits that form a credit card number) recognition, the words are surrounded by pauses (silence). This type of recognition is easier than continuous speech recognition because, in the latter, the word boundaries are not so evident. In addition, the level of difficulty varies among the continuous speech recognition due to the type of interaction. That is, recognizing speech from human-human interactions (recognition of conversational telephone speech, broadcast news) is more difficult than human-machine interactions (dictation software) [8]. In read speech or when humans interact with machines, the produced speech is simplified (slow speaking rate and well-articulated) so that it is easy to understand it [7].

Speaker mode: the recognition system can be used by a specific speaker (speaker dependent) or by any speaker (speaker independent). Even though speaker dependent systems require to be trained on the user, they generally achieve better recognition results (there is no much variability caused by the different speakers). Given that speaker independent systems are more appealing than speaker dependent ones (no training required for the user), some speaker-independent ASR systems are performing some type of adaptation to the individual user ‘s voice to improve their recognition performance.

Channel type: the characteristics of the channel can affect the speech signal. It may range from telephone channels (with a bandwidth about 3.4 kHz) to wireless channels with fading and with a sophisticated voice [6].

Transducer type: defines the type of device used to record the speech. The recording may range from high-quality microphones to telephones (landline) to cell phones to array microphones (used in applications that track the speaker location).

Fig. 4 shows the progress of spoken language systems along the dimensions of speaking style and vocabulary size. Note that the complexity of the system grows from the bottom left corner up to the top right corner. The bars separate the applications that can and cannot be supported by speech technology for viable deployment in the corresponding time frame.



Progress of spoken language system along the dimension of speaking style and vocabulary size (adapted from [9]).

Some other parameters specific to the methods employed in the development of an ASR system are going to be analyzed throughout the text.

2.1.2 Text-To-Speech

Flite Speech Synthesis Engine: Flite is an open source, fast run-time text to speech synthesis engine developed at CMU, which is primarily designed for small embedded devices [23]. Built speech synthesis or the festvox voice using Festival Speech Synthesizer will be converted to flite voice to use in android platform. Festvox voice contains the natural language processing unit and digital signal processing unit. In Text-To-Speech synthesizer text analysis and linguistic analysis modules together are known as Natural Language Processing unit. Also waveform generation is known as Digital Signal processing unit. There are three main waveform generation technologies.

- Formant Synthesis

In this technique speech output is created using an acoustic model. Parameters such as fundamental frequency, noise levels are varied over time to create a waveform of artificial/Robotic speech. [24] This generates acoustic structure by rule.

- Concatenative Synthesis

This technique gives the most natural sound in speech synthesis. Concatenative speech synthesis is based on the idea of concatenating pre-recorded sound units to construct the utterance. There are two approaches in this.

I. Diphone Synthesis - Diphone synthesis generates the speech waveform by concatenating basic speech segments name diphones. A diphone is recording of the transition of two adjacent pair of phones. [25] This produce very intelligible synthetic speech. UCSC followed this technique to create their system.

II. Unit Selection Synthesis - This approach reduces the fixed-size unit synthesis. The speech output is produced by selecting and concatenating sounds or words from a speech database. The corpus database is searched for maximally long phonetic strings to match the sounds to be synthesized. The quality rises because of the number of concatenation points decreases.

- **Statistical-parametric speech synthesis**

Statistical-parametric speech synthesis can use with Hidden Markov Model (HMM). The models trained on speech corpora and no data needed at runtime.

The voices was build using the Clustergen [26] method of statistical parametric synthesis within the Festival/ FestVox voice building environment.

Following are the UML diagrams for the better understanding of the entire architecture of the application:

- Class Diagram for Sanwadha : Refer Appendix B: figure B.1
- Use Case Diagram for Sanwadha : Refer Appendix B: figure B.2