A Sophisticated HCI Perspective: Advanced Bengali Phonetics Communication System for Disabled (Deaf and Dumb) Persons

Nazmus Sakib, Chowdhury Amlan, Mohaiminul Islam, Maisha Maliha, Salowa Samin
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology, Dhaka, Bangladesh

ABSTRACT

The purpose of this study is to develop an intelligent system that can help the people who are dumb and deaf as they are recognized as disabled persons. These people face difficulties in communicating with other people in their regular life activities such as travelling, shopping etc. Nowadays, Human Computer Interaction (HCI) domain facing and solving the problem of disabled persons of this day to day activities. In this paper a new solution of this problem and the perspective of HCI is being focused. This paper is synthesized and analyzed speech-to-text (STT) and text-to-speech (TTS) algorithms to implement the system for those people who are deaf & dumb. The system will behave and act on behalf of the person and it will also capture the words from the other side and convert it to text, thus the deaf person may understand others say through the display of this system. There is a huge amount of people who are deaf and dumb in Bangladesh. Therefore, the system has been implemented both in Bengali phonetics & English language. The performance of this paper showed a satisfactory result of this system.

Keywords

Speech-to-text (STT), Text-to-speech (TTS), Deaf People, Dumb People, Handicapped people, Human Computer Interaction (HCI), Disabled People, Sphinx4, MaryTTS, RTRL.

1. INTRODUCTION

Nowadays, humans are blessed with the human computer interaction which is a study that interacts with the computers and the extents of computers for the development of human behavior for successful interaction. During the last three decades HCI has expanded firmly and rapidly by attracting professionals from various disciplines and organizing diverse concepts and approaches [1]. HCI now aggregates a collection of semi-autonomous fields of research and practice in human-centered informatics. HCI is essential in some case of making any ergonomic product which is more useful, functional, successful and safe. Ultimately it is all about developing a pleasurable and attractive product for the users without distracting them from their tasks [2].

Human Interaction is the way of how people change the environment and vice versa. It concerns about the uses of computer and communication system by the people as individual or as groups and also affected by the devices. It also helps in improving the interaction between a person and a computerized device in such an effective way that the person is benefitted immensely. It's all about the study and design planning of the way how human and a computer network work together. [2]

Three main segments of HCI consists of- the users, the computer itself and the ways they are working together. Firstly, the users are those people who want to get their works done by using technology. It may be an individual user or a group of users of an organization. Secondly, the computer refers any process control system or desktop or large scale of computer systems. And last one is the interaction which can be done by user and computer. It's the attachment between which is required by the user and what can be done by the computer. [2]

Currently people are trying to use HCI in each and every possible sectors of our daily life. Here this study represents a communicator system for disabled people. It helps to build an advanced duplex (both way communication) communicator system with an attractive and user friendly User Interface (UI) which will help disable people who are deaf and dump, for their regular communication. As the system is implemented now for desktop computers but in extent the research will be implemented for android and iOS. The prime motive is to formulate a system which would give optimum performance in terms of complexity, accuracy, delay and memory requirements for mobile environment [3].

The mechanism has been used mainly the STT algorithm and TTS algorithm by adding phonetic a dictionary [4] [5].These algorithms are both complex phenomenon. We have studied a lot of algorithms for STT, TTS and also some noise reducing filters. There are some STT filters. But we have used sphinx-4 that uses Hidden Markov Model (HMM) which is for speech to text conversion. Sphinx4 is a pure Java speech recognition library [4]. And MaryTTS for text to speech conversion as their performance is much satisfactory among all [6].

The rest of this paper is organized as follows. Sections 2 and 3 describe the Literature Review and Proposed Methodology. Section 4 and 5 provides details of implementation and the result analysis. Finally, section 6 draws the conclusion of this paper with a few comments and suggestions on future research.

2. LITERATURE REVIEW

For this research many of the techniques are being reviewed. For the development of this research some verified topic need to be analyzed.

2.1. Speech-To-Text (STT)

Speech–to–text research has found new idea to help the handicap people with the voice prompted writing tools [7]. Speech is the most frequent but complex phenomenon which is produced and perceived in a complex way. The naive perception is often that it is built with words, and each word consists of phones. Speech is, however, a dynamic process without clearly distinguished parts. It converts the spoken words into text form exactly in the similar way that the user pronounces [8].All modern descriptions of speech are to some degree probabilistic. That means that there are no certain boundaries between units, or between words. Speech to text translation and other applications of speech are never been 100% correct. It works like a continuous audio stream where stable states may mix with dynamically unstable states [8].

A conversion method can be developed from speech to text using SAPI for Bangla language [9]. SAPI is Speech Application Program Interface developed by Microsoft Corporation which is used for speech related works in windows operating system. But it works only for eight languages. Hence English language was used for speech to text conversion in SAPI for Bangla. SAPI returns Bangla words in English if matches work. Then the matched word fetched the Bangla words form the database. The experimental recognition rate was nearly 78% on average [9].

Focusing on the system speech to text for deaf and hard-of-hearing people, a word sound or phoneme can be used as a palantype based systems [10]. The system will have a base dictionary which consist English words linked to Palanforms [11]. What is actually necessary for deaf people – this was the actual target. The subtitling concept was used here which needs to be done live [10].

There are some scientific problems that create a hazard to achieve the basic goals of the STT system which is getting the human-sounding speech. It was done on linguistic frame work too. Many examples are taken from the Klattalk algorithm to develop the system which drives a format synthesizer, articulatory synthesis and waveform concatenation [12].There is an efficient system which performs the speech to text conversion based on Spanish language. It carried out a clean output on Spanish. At the end of the experiment the whole speech-to-text system neatly outperforms the word-constrained baseline system. The system was developed on Spanish which continuously receives a source of alphanumeric characters [13].

For speech to text conversion a synthetic speech can also be used by vocally handicapped person. Then it can be a source of learning for visually impaired person. It can also be used in games and education, in telecommunication and multimedia and for a voice enabled email. Automatic speech recognition system is also described in this paper. The basic principles was discussed and the extraction of the features of it too [14].

In this system, an API library named Sphinx-4 that uses Hidden Markov Model (HMM). Sphinx4 is a pure Java speech recognition library. It provides a quick and easy API to convert the speech recordings into text with the help CMUSphinx acoustic models. Beside speech recognition Sphinx4 helps to identify speakers, adapt models, align existing transcription to audio for times tamping and more. As we are going to need a large vocabulary decoder, there must be diarization framework, adaptation framework and post processing framework. They are all needed to be cooperated by somehow. Flexibility of sphinx4 will allow us to build such a system quickly. It's easy to manage many sphinx4 instances doing large-scale decoding on a cluster. So, because of so much adjustability and modifiable capacity, we have chosen to use Sphinx-4 Library [4].

The Knowledge Base (KB) or Linguist provides the information the decoder needs to do its job. It is made up of three modules: Acoustic Model, Dictionary and Language Model [15]. Acoustic Model contains a representation (often statistical) of a sound, created by training using acoustic data. Dictionary has the pronunciation of all the words to be recognized. Language Model contains a representation (often statistical) of the probability of occurrence of words. The decoder is the main block of Sphinx-4 and performs the actual recognition. It comprises a graph construction module, which translates any type of standard language model provided to the KB by the application into an internal format, and together with information from the dictionary, and structural information from one or more sets of acoustic models, constructs a Language HMM [16] [4].

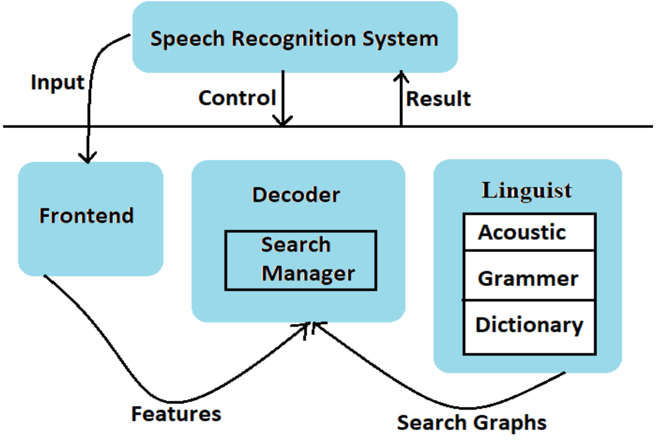


Fig. 1. Sphinx-4 configuration reading speech recognition

2.2. Text-To-Speech (TTS)

Developing text-to-speech capability includes some unique challenges. Especially in the English language, where a great number of homonyms have varied pronunciations, computer programs rely on probability modeling to guess the desired pronunciation of a word in digital text. The program also has to convert units of text into phonemes, the smallest units of speech pronunciation [17]. Basic Steps of Text-to-Speech:

i. Text analysis
ii. Automatic phonetization
iii. Prosody generation

There are some more researches on Bangla text to speech conversion. It is a simple phoneme conversion system. A method for producing natural speech sound during the conversion of speech to text was developed. Creating a dataset by recording all the Bangla alphabet sound, which will be used for the conversion of input text into corresponding speech. This system supports UNICODE inputs for Bangla text. Though distortion occurs in this system, it is easy to use and is an efficient system [18].

A recurrent neural networks trained with the Real Time Recurrent Learning (RTRL) algorithm can be used too for text-to-phoneme conversion [19].The real RTRL algorithm actually takes a long time. It does so because their computational complexity is high. So a fast RTRL algorithm which computational complexity is comparatively lower was proposed and it really help in fast learning. [20] The recurrent neural networks focused on the spatial temporal problems such as an area of artificial intelligence which draws from the fields of computer science, cognitive science, and cognitive psychology [21]. This kind of problem was handled in converting English text streams to speech. For n number of processing nodes, the computational complexity is O(n4). Here, n is the number of the nodes which will be processed in the output layer. But they are stating that their system aims at speeding up the training phase while maintaining the whole feature of RTRL algorithm. By connecting link between input & output layer, the system confines the error back propagation to some randomly selected connections [19].

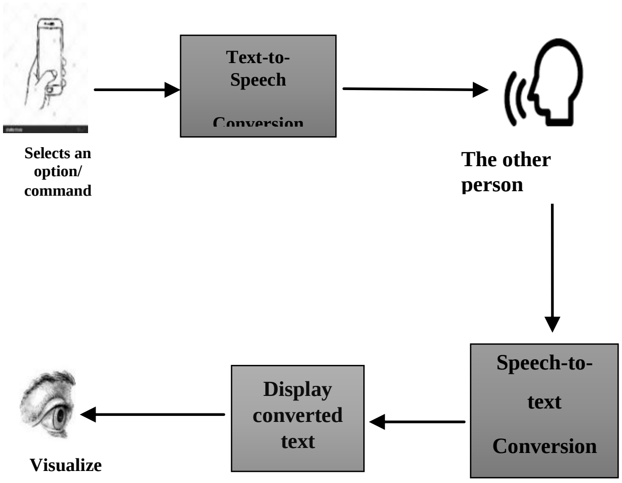

Text-to-Speech Conversion
Selects an option/ command
The other person
Speech-to-text Conversion
Display converted text
Visualize

Fig. 2.Model for both Deaf and Dumb person

A research on Bangla text to speech conversion, in which recorded audio sound of the Bangla phonemes and syllables was used where the system searches for the best matches of syllables and then phonemes. No phoneme based method is needed. Syllable based methods gives a better quality speech for the input text [22]. Implementation of text-to-speech in effective way, we will use MaryTTS which is an open-source, multilingual Text-to-Speech Synthesis platform. The reason of using MaryTTS is that it comes with toolkits for quickly adding support for new languages and for building unit selection and HMM-based synthesis voices [5] [6].

2.3. Speech Recognition

Type of Speech recognition system can be separated in different classes by describing what type of utterances they can recognize.

2.3.1. Isolated Word

Isolated word recognizes attain usually require each utterance to have quiet on both side of sample windows. It accepts single words or single utterances at a time. This is having “Listen and Non Listen state”. Isolated utterance might be better name of this class.

2.3.2. Connected Word

Connected word system is similar to isolated words but allow separate utterance to be “run together minimum pause between them.

2.3.3. Continuous speech:

Continuous speech recognizers allows user to speak almost naturally, while the computer determine the content. Recognizer with continues speech capabilities are some of the most difficult to create because they utilize special method to determine utterance boundaries.

Spontaneous speech: At a basic level, it can be thought of as speech that is natural sounding and not rehearsed an ASR System with spontaneous speech ability should be able to handle a variety of natural speech feature such as words being run together. A real time speech recognition system is tested in real time noisy environment. The bidirectional non stationary Kalman filter is used to enhance the ability of Real time speech recognition system. Bidirectional Kalman filter has been proved to be the best noise estimator in non-stationary noisy environment [23] [24].

The development of an efficient speech recognition system can be done by using different techniques such as Mel Frequency Cestrum Coefficients (MFCC), Vector Quantization (VQ) and Hidden Markov Model (HMM) [25]. Speaker recognition followed by speech recognition can be used to recognize the speech faster, efficiently and accurately. MFCC is used to extract the characteristics from the input speech signal with respect to a particular word uttered by a particular speaker. Then HMM is used on Quantized feature vectors to identify the word by evaluating the maximum log likelihood values for the spoken word. In the speaker identification phase, MFCC and Distance Minimum techniques have been used. These two techniques provided more efficient speaker identification system. The speech recognition phase uses the most efficient HMM Algorithm. It is found that Speaker recognition module improves the efficiency of speech recognition scores. It has been found that the combination of MFCC and Distance Minimum algorithm gives the best performance and also accurate results in most of the cases with an overall efficiency of 95%. The study also reveals that the HMM algorithm is able to identify the most commonly used isolated word. [26]

Research in Bengali speech recognition field is still in primary stage. A large amount of progressive work are required in the field of Bengali speech–to–text conversion. [27]

2.3.4. Speaker Model

Types of Speaker Model: All speakers have their special voices, due to their unique physical body and personality. Speech recognition system is broadly classified into main categories based on speaker models, namely, speaker dependent and speaker independent.

3. PROPOSED METHODOLOGY

A disabled person who is deaf and dumb can communicate socially without hassle in their everyday life through this interactive system. This system will basically consist of speech-to-text and text-to-speech, that is, a both way communication process in Bengali phonetics and also in English.

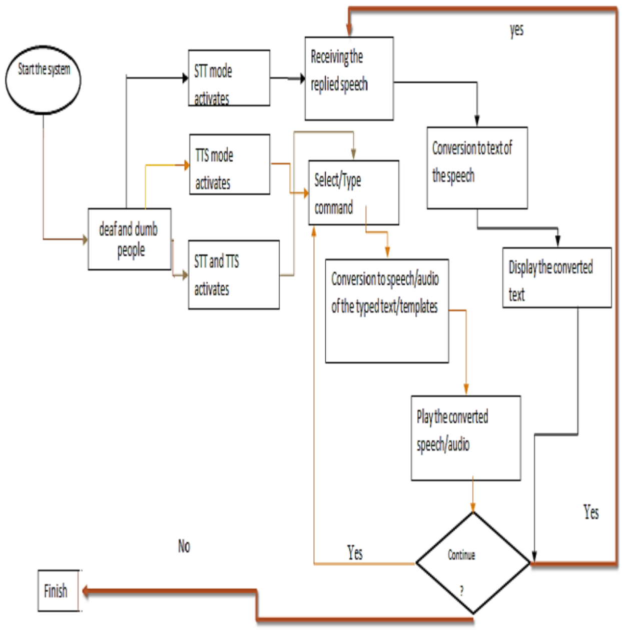

Start the system
STT mode activates
Receiving the replied speech
yes
TTS mode activates
Select/Type command
Conversion to text of the speech
deaf and dumb people
STT and TTS activates
Conversion to speech/audio of the typed text/templates
Display the converted text
Play the converted speech/audio
Yes
No
Yes
Continue ?
Finish

Fig. 3.flow diagram for proposed model

Firstly, the system has been done in duplex mode as (i) Speaking (STT) and (ii) Listening (TTS). The person can choose either STT or TTS whatever they need. The system will be specialized for use in these purposes:

i. Transportation

ii. Shopping

iii. Restaurant

As our Bangladeshi English accent is not same as the British accent, so it may be difficult for system to recognize accent quickly. To improve recognition, the system will have to adapt acoustic model. The adaptation of acoustic model improves the fit between the adaptation data and the model. The state will perform cross language adaptation and adapt English model to sounds of other language by creating a phone set map. The adaptation process takes transcribed data and improves the model of system model. For adaptation process Adapting Acoustic Model has been used. It follows some steps-

i. Creating an adaptation corpus
ii. Recording adaptation data
iii. Generating acoustic feature files
iv. Converting the send ump and mdef files
v. Accumulating observation counts
vi. Creating transformation with MLLR
vii. Updating acoustic model files with MAP
viii. Recreating the adapted send ump file

As disabled person will use our system in real time based situation, it need to make sure that speech recognition process is smooth enough. To make speech recognition more accurate and fast, this paper is going to state on two main factors, those are Phonetic Dictionary and Acoustic Model.

3.1. Text-to-speech Implementation

A Text-To-Speech (TTS) synthesizer is a computer-based system that should be able to read any text aloud, whether it was directly introduced in the computer by an operator or scanned and submitted to an Optical Character Recognition (OCR) system. Let us try to be clear. There is a fundamental difference between the system about to discuss here and any other talking machine (as a cassette-player for example) in the sense that we are interested in the automatic production of new sentences. This definition still needs some refinements. Systems that simply concatenate isolated words or parts of sentences, denoted as Voice Response Systems, are only applicable when a limited vocabulary is required (typically a few one hundreds of words), and when the sentences to be pronounced respect a very restricted structure, as is the case for the announcement of arrivals in train stations for instance. In the context of TTS synthesis, it is impossible (and luckily useless) to record and store all the words of the language. It is thus more suitable to define Text-To-Speech as the automatic production of speech, through a grapheme-to-phoneme transcription of the sentences to utter.

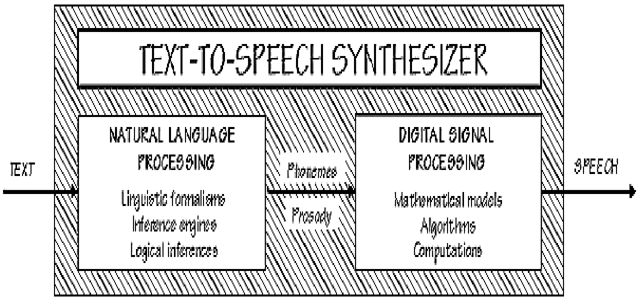


Fig. 4.Basic Text-to-Speech Synthesizer [28]

Unit-Selection Voice building has some prerequisites as All speech signals are in .wav format, 16 kHz sampling frequency and 16-bit sample format. Next, all speech signals have corresponding text of what is spoken in each speech signal (.txt format) and MaryTTS server must be running on localhost:59125(default).

4. IMPLEMENTATION

The system has been implemented as a desktop application using JAVA because it's able to run on multiple platforms using its own run time environment JVM (Java Virtual Machine). Here, eclipse is been used as our programming environment. It's available for most of the operating systems. Moreover, CMU is used Sphinx for speech to text conversion and MaryTTS for text to speech conversion. For speech recognition a phonetic dictionary is used which is added to the JRE library and different grammar file (.gram file) for identifying local language in English letter.

The development of text to speech using MaryTTS which uses three ways to identify a word- acoustic model, dictionary and grammar file. Phonetic dictionary provides system the data to map vocabulary words to sequence of phonemes. The phonetic dictionary looks like this. For example:

```
hello H EH L OW
world W ER L D
```

Because the dictionary already contains the phonetics for traditional English words, initially, new words are not needed to add, but this can see in the above purposes, if this system is made for the people of Bangladesh, lots of words out of traditional English dictionary has to be added. I.e. for transportation purpose – Rickshaw, Tom-Tom etc.; for restaurant purpose – Chotpoti, Fuchka etc; for shopping purpose- Sharee, Punjabi, Salowar kamiz etc. and other general purposed also. Therefore the phonetic dictionary will be edited with a view to increasing the conversion accuracy. In order to extend an existing dictionary for new words we are going to use G2P-seq2seq. It is based on neural networks implemented in Tensorflow framework and provides a state of the art accuracy of conversion. The output of the system is as shown below. At first, one of the modes has to be chosen among three modes. Then, user can use it either in Bengali or English for speaking & listening:

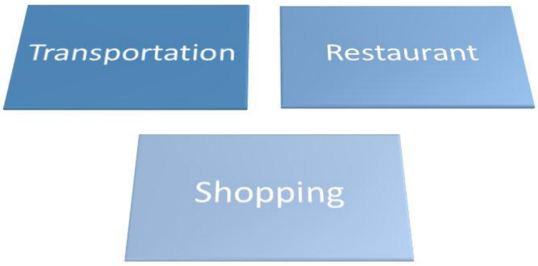
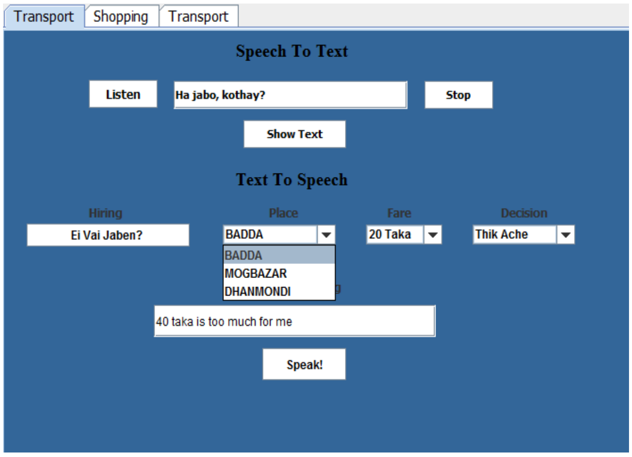


Fig. 5. Three modes



Speech To Text

Ha jabo, kothay?

Text To Speech

Hiring: Ei Vai Jaben?

Place: BADDA (BADDA, MOGBAZAR, DHANMONDI)

Fare: 20 Taka

Decision: Thik Ache

40 taka is too much for me

Fig. 6. Output for both Bengali & English

5. RESULT ANALYSIS AND DISCUSSION

For speech to text result analysis here ten different types of speaker's accent is used. A list of words were given to the speaker for to test the accuracy of our system. This process has gone through a lab test where people acted as a deaf & dumb through the process.

$$\text{Accuracy (y)} = f(x) = \frac{d}{dw}\int_{n=100}^{n=1}\frac{w_r}{w} \times 100\%$$

w_r = number of recognized words
w = total number of spoken words

Accuracy rate of our system was nearly 75.84%.
The test has been taken place for 3 days. Category 1 is for 1st day when the people used the system for the 1st time for the three environment (transportation, shopping & restaurant). Category 2 is for 2nd day. Category 3 is for 3rd day when they knew how to handle the System. They were more comfortable than the 1st day to use the system.

For three days the same persons used the system. And tremendous result has seen for the second and the third day. The accuracy level for 2nd and 3rd day jumped up to 69.4% for sector 2 and 84.9% successively. Once users get used to with the system, they are comfortable to use it gradually.

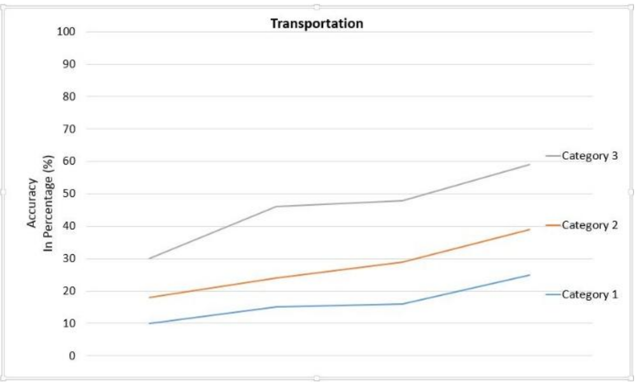


Fig 7: accuracy level for Transportation

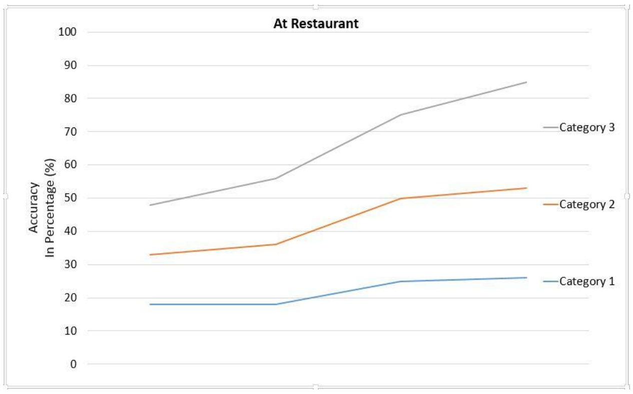


Fig 8: Accuracy level for Restaurant

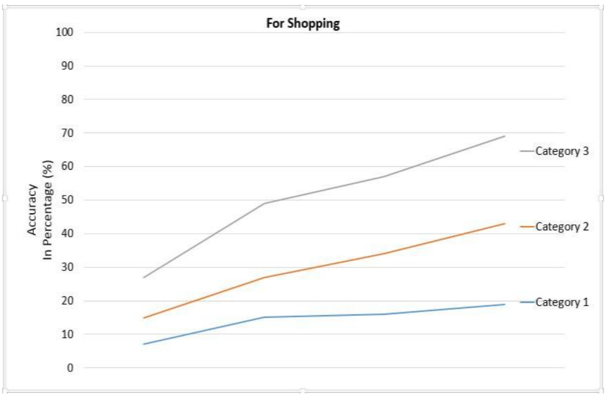


Fig 9: accuracy level for ordering at shopping

Review has been taken from the people how they think that the system is useful. We have a set of questionnaire for them:

a. How is the system? (Good/Satisfactory/Dissatisfactory)
b. Do you think the system is useful for deaf & dumb people? (Yes/No)
c. Do you think the system is suitable for kids in school? (Yes/No)
d. How will you rate this system out of 5? (1/2/3/4/5)

People rated our system out of 10 against these bunch of questions-

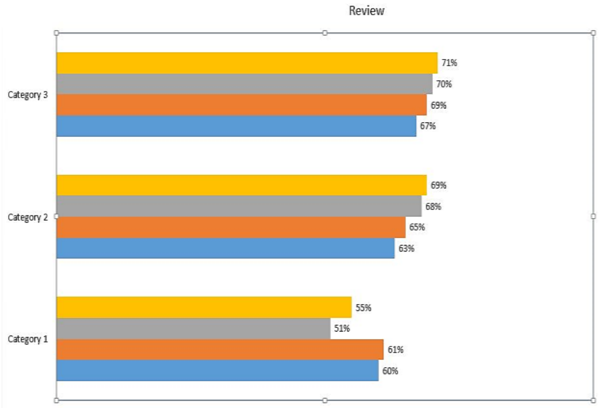


Fig 10: review out of 10

The review has been categorized in percentage with the respect of three days' testing. On the third day a good review has been goten by the users.

6. CONCLUSION

It has been suggested that one of the most promising areas for the application of speech recognition is in helping handicapped people. Within the last decade, many improvements have been made in the performance of automatic speech recognizers and current technology is discussed in relation to the needs of the disabled population. In this report, how a deaf and dumb people can be helped in various ways through this system is discussed. It's a comprehensive communicator for the specific type of disabled people.

7. FUTURE PLAN

We are working on speech-to-text and text-to-speech conversion. In future we want to make this process more efficient as much as possible and we will be trying to show the test in Bengali letters for android and iOS. Hope it will also enable a user to perform operations such as open calculator, WordPad, notepad, log off computer, especially for those who are unable to use keyboard. We will try to make this more accurate, reduce errors, noise and completion time. The accuracy of speech recognition varies with vocabulary size, confusability and isolated, discontinuous, or continuous speech. So we are trying to improve these. We will also to try to design the microphone and sound system more efficient that our system will be able to adapt more quickly.

8. REFERENCES

[1] "Macintosh human interface guidelines," 1992. [Online]. Available: https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/human-computer-interaction-brief-intro. [Accessed 2016].

[2] "Human-Computer Interaction and Your Site SitePoint," [Online]. Available: https://www.sitepoint.com/computer-interaction-site/. [Accessed 2016].

[3] R. Sandanalakshmi, P. A. Viji, M. Kiruthiga, M. Manjari and A. Sharina, "Speaker Independent Continuous Speech to Text Converter for Mobile Application".

[4] "CMUSphinx Tutorial For Developers [CMUSphinx Wiki]," [Online]. Available: http://cmusphinx.sourceforge.net/wiki/tutorial. [Accessed 14 03 2017].

[5] "MaryTTS – Introduction," [Online]. Available: http://mary.dfki.de/. [Accessed 2016].

[6] "MaryTTS – Project Information," [Online]. Available: http://mary.dfki.de/project-info.html. [Accessed 2016].

[7] A. V. Bapat and L. K. Nagalkar, "Phonetic Speech Analysis for Speech to Text Conversion," 2008 IEEE Region 10 and the Third international Conference on Industrial and Information Systems, pp. 1-4, 2008.

[8] S. K. Gaikwad, B. W. Gawali and P. Yannawar, "A Review on Speech Recognition Technique," International Journal of Computer Applications, vol. 10, no. 3, pp. 0975-8887, 2010.

[9] S. Sultana, M. A. H. Akhand, P. K. Das and M. M. H. Rahman, "Bangla Speech-to-Text conversion using SAP," 2012 International Conference on Computer and Communication Engineering (ICCCE), pp. 385-390, 2012.

[10] C. Brookes, "Speech-to-text systems for deaf, deafened and hard-of-hearing people," Speech and Language Processing for Disabled and Elderly People (Ref. No. 2000/025), IEE Seminar on, no. May, pp. 5/1-5/4, 2000.

[11] "Speech-to-text reporter - Wikipedia," [Online]. Available: https://en.wikipedia.org/wiki/Speech-to-text_reporter. [Accessed 2016].

[12] D. H. Klatt, "Text-to-speech Conversion".

[13] M. Penagarikano and G. Bordel, "Speech-to-text translation by a non-word lexical unit based system," Signal Processing and Its Applications, vol. 1, 1999.

[14] P. Khilari and P. V. P. Bhope, "Implementation of Speech-to-Text Conversion," vol. 4, no. 7, 2015.

[15] W. Pedryez, "Linguistic models and linguistic modeling," IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics), vol. 29, no. 6, pp.

745-757, 1999.

[16] "Natural Language Processing and Information Systems: 15th International ... - Google Books," [Online]. Available:

[17] "An Introduction to text-to-speech synthesis," [Online]. Available: http://tcts.fpms.ac.be/synthesis/introtts_old.html. [Accessed 2016].

[18] S. #. b. ". M and H. h. i. a. *QMp2`bBQM, Qbb, D M J X H, pp. 6-8.

[19] Y.-L. Lu, Mak, Man-Wai and W.-C. Siu, "Application of a fast real time recurrent learning algorithm to text-to-phoneme conversion," IEEE International Conference on Neural Networks, 1995. Proceedings, vol. 5, pp. 2853-2857, 1995.

[20] "Deep Voice: Real-Time Neural Text-to-Speech for Production - Baidu Research," [Online]. Available: http://research.baidu.com/deep-voice-production-quality-text-speech-system-constructed-entirely-deep-neural-networks/. [Accessed 2016].

[21] "Spatial Temporal Patterns for Action-Oriented Perception in Roving Robots - Google Books," [Online]. Available:

[22] M. Y. Arafat, "Speech synthesis for Bangla Text to Speech conversion," SKIMA 2014 - 8th International Conference on Software, Knowledge, Information Management and Applications, pp. 0-5, 2014.

[23] "Kalman filter toolbox for Matlab," [Online]. Available: http://www.cs.ubc.ca/~murphyk/Software/Kalman/kalman.html.

[24] "A real time speech to text conversion system using bidirectional Kalman filter in Matlab," [Online]. Available: https://www.semanticscholar.org/paper/A-real-time-speech-to-text-conversion-system-using-Sharma-Sardana/592c72ee51f774d47b35528e0512dc9b02000277.

[25] "An Efficient Speech Recognition System," [Online]. Available: https://www.researchgate.net/publication/263358996_An_Efficient_Speech_Recognition_System.

[26] "What is a hidden Markov model?," [Online]. Available: https://www.nature.com/nbt/journal/v22/n10/full/nbt1004-1315.html.

[27] "An Introduction to text-to-speech synthesis," [Online]. Available: http://tcts.fpms.ac.be/synthesis/introtts_old.html. [Accessed 25 10 2016].