Group No. 6

Section 2

Instructor: Dr Sayantan Chakraborty

Group Members	ID Number
1. DEEKSHA SACHAN	2020A2PS0001U
2. AAYUSH KAPOOR	2020A7PS0009U
3. MEGHA MANOJ	2020A7PS0016U
4. AAMINA TASKEEN	2020A7PS0025U
5. AYUSH MALLICK	2020A7PS0257U
6. GLADWIN PAUL	2020A7PS0258U

Birla Institute of Technology and Science

D.I.A.C, Dubai, UAE

A Report

on

AUTOMATIC SPEECH RECOGNITION

Prepared for

Dr Sayantan Chakraborty

Instructor in charge

By

Group 6

Approved by

Dr Sayantan Chakraborty

Instructor in charge

March 2021

ABSTRACT

Our research identifies the problems that are associated with ASR systems and explains how the efficiency of an ASR system can be improved. We have surveyed a sample group to investigate the awareness and diverse capabilities of the ASR technology. The problems that surfaced are 30% error in accuracy levels, rare acceptance of different accents, and occurrence of background noise while using ASR. These problems can be tackled by decreasing the error rate, including different accents by getting more audio inputs, and improving the quality of microphones to reduce background noises. People with speech impairments have been excluded and hence ASR platforms need to diversify their user base. ASR has outstanding potential and with further development can be revolutionary.

ACKNOWLEDGEMENTS

We are tremendously grateful to Dr. Shazi Shah Jabeen, Instructor in Charge, and Dr. Sayantan Chakraborty for this wonderful opportunity to explore the topic Automatic Speech Recognition (ASR), an interdisciplinary subfield of computer science. This project has been extremely helpful in understanding the problems faced with this technology.

We would also like to express our gratitude to, Technical Report Writing Faculty, for clearing our queries and assisting us in creating the Turnitin account.

We would also like to thank the surveyors for taking out their time to fill up the survey and providing us with insight into the ASR field.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1. AUTHORIZATION

The present report based on automatic speech recognition was approved and authorized by Sayantan Chakraborty, instructor in charge, BITS, UAE on 24 march 2021.

1.2. HISTORICAL BACKGROUND

The humankinds' interest in using speech as a technology led to the formation of the automated speech recognition such as Alexa, Cortana, Google, and Siri, otherwise known as ASR. Since its first design in the 1950s as Audrey, it has been under extensive research to improve its accuracy and functioning.

However, it has not been able to achieve its full potential in terms of its accuracy. The fact that it is unable to grasp the words of the user is the main issue. This may be due to the accent of the individual, the environment in which it is used, or the inability of the person to communicate properly due to physical conditions. To address such issues a survey was conducted with the help of which the report was formed and the recommendations were made.

1.3. OBJECTIVES

This project has been undertaken with the intent of understanding the technical problems faced by ASR. The information obtained through a survey has been used to:

- To highlight the technical issues faced by ASR technology.
- To attain extensive knowledge on the identified technical issues with reference to the survey data.
- To propose alternate methods for improvement in the accuracy levels.

1.4. SCOPE

- The survey was conducted in Dubai, Kuwait and India.

- The survey was conducted for all age groups under 60.

1.5. LIMITATIONS

- As mentioned above the report was limited to Dubai, Kuwait and India, thus the data displays the utilization of speech recognition systems only in these places. Other places were not covered due to time constraint.

- The survey was restricted to people under 60, thus the data was collected from teenagers and adults. Elderly people above the age of 60 were not considered due to inadequate resources.

1.6. METHODS AND SOURCES OF DATA COLLECTION

The data for this report was collected using a survey by questionnaire. A questionnaire containing seven questions was distributed through the mail to 500 people residing in the above-mentioned countries. All the residents were randomly selected and belonged to the age groups: below 18, 18-30, and 30-60. Out of 500, only 200 people returned the filled questionnaires. In addition to this, 30 people were interviewed, that is, 6 from each country.

1.7. REPORT PREVIEW

Besides Introduction, the report contains three chapters. Chapter 2 gives an analysis of the usage of ASR and the problems they face while using it based on statistical data. Chapter 3 sums up the discussion whereas Chapter 4 contains suggestions and recommendations.

CHAPTER 2: DISCUSSION

2.1. ACCURACY LEVEL

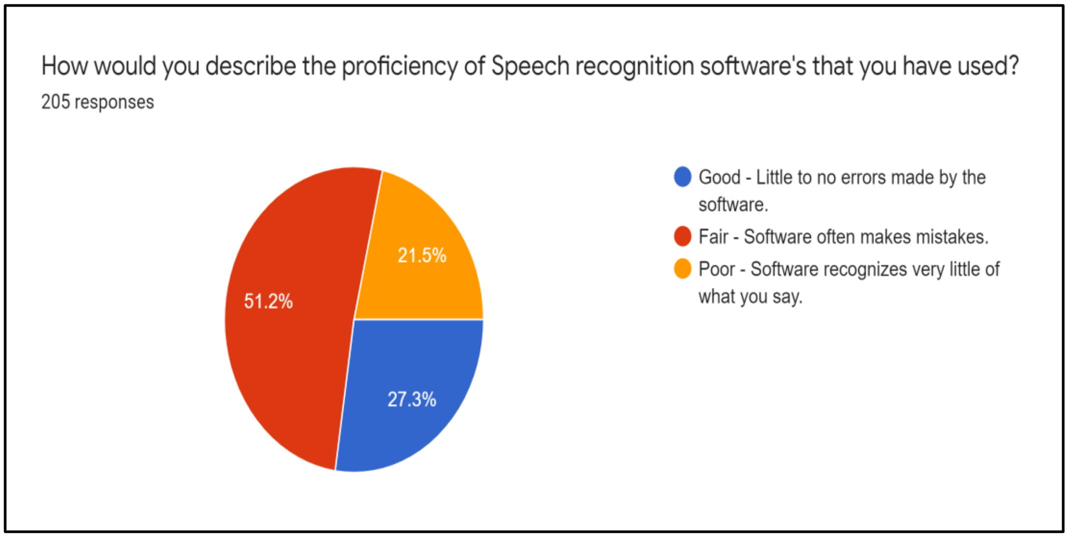
2.1.1. WORD ERROR RATES

Word error rate (WER) is the sum of incorrect words identified during recognition (I), words that are undetected (D), words that are substituted (S), then divided by the total number of words provided in the human-labeled transcript (N). Finally, that number is multiplied by 100% to calculate the WER.

$$WER = \frac{I + D + S}{N} * 100\%$$

Figure 2.1.1 Formula for calculating WER.

Mostly ASR has about 30% WER and causes many issues like omitting words, and weighing all the words equally.



How would you describe the proficiency of Speech recognition software's that you have used?

205 responses

Figure 2.1.2 Proficiency of ASR that is used on daily bases.

2.1.1. VOCABULARY JARGON AND HOMOPHONES

The software is never able to have a complete list of all the words that are present in any language. Words like jargons that have very specific meanings in specific situations couldn't completely be listed. Homophones that sound very similar are very problematic when it comes to being processed. This is due to the inability of the software to have a contextual relationship between words like that of humans.

2.1.2. FLEXIBILITY OF DETECTION

The most common approach to designing ASR Systems is to train acoustic models to analyze and recognize small pieces of audio called phonemes. The ASR decoder takes each phoneme and compares it against all possible phonemes. This process has to be repeated hundreds of times for each sentence, making it highly inefficient and slow. Furthermore, the performance of speech recognition systems drastically worsens when users use new words and slangs, when two or more people speak simultaneously or when users combine words from two languages as a force of habit.

2.2. ACCENT AND DISABILITIES

2.2.1. SPEECH IMPAIRMENT

Speech recognition up to some extent has helped those with handicaps among its wide range of users. However, it has yet to be proven useful for those with speech impairments such as stutter, down’s syndrome, etc.

While about 60% of the disabled population worldwide have difficulties communicating orally with other people, the level of hardship increases while interacting with computerized technology. About 9 out of 205 people surveyed have trouble using ASR due to intra-word pauses, stutter, breathing sounds, and unclear speech that are aspects of their vocal impairments. The system's inability to understand what they say due to these conditions make its application inefficient for them.

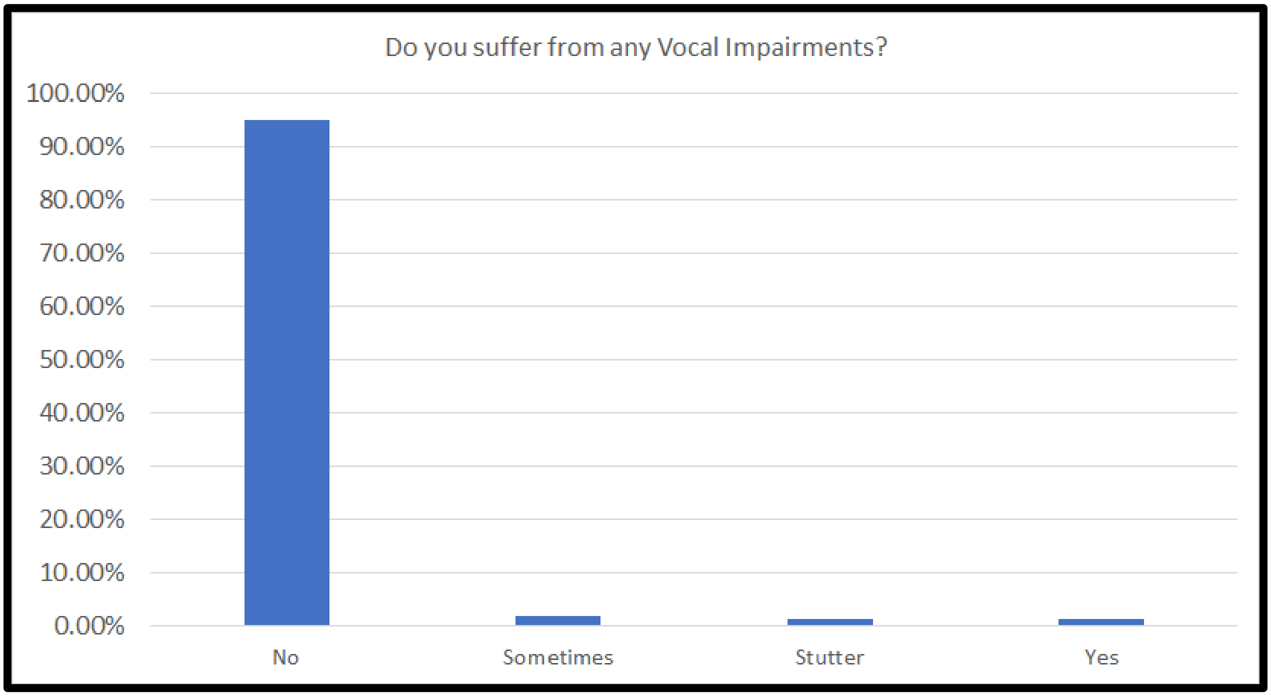


Figure 2.2.1. ASR users with vocal impairments.

2.2.2. LACK OF DIVERSITY

Since the introduction of speech recognition, the default language recognized by the system is English with western accents. Even after decades of development, the ASR system fails to recognize different accents and does not work in numerous languages.

23.5% of the survey responses reveal that the software does not recognize what they say due to their accent while 44% encounter this issue occasionally. The difference in pitches, tones of speech, and different pronunciations vary with regions. The lack of data procured in terms of phonetic transcription, the representation of speech sounds utilizing symbols, for interpretation plays a major role in this problem.

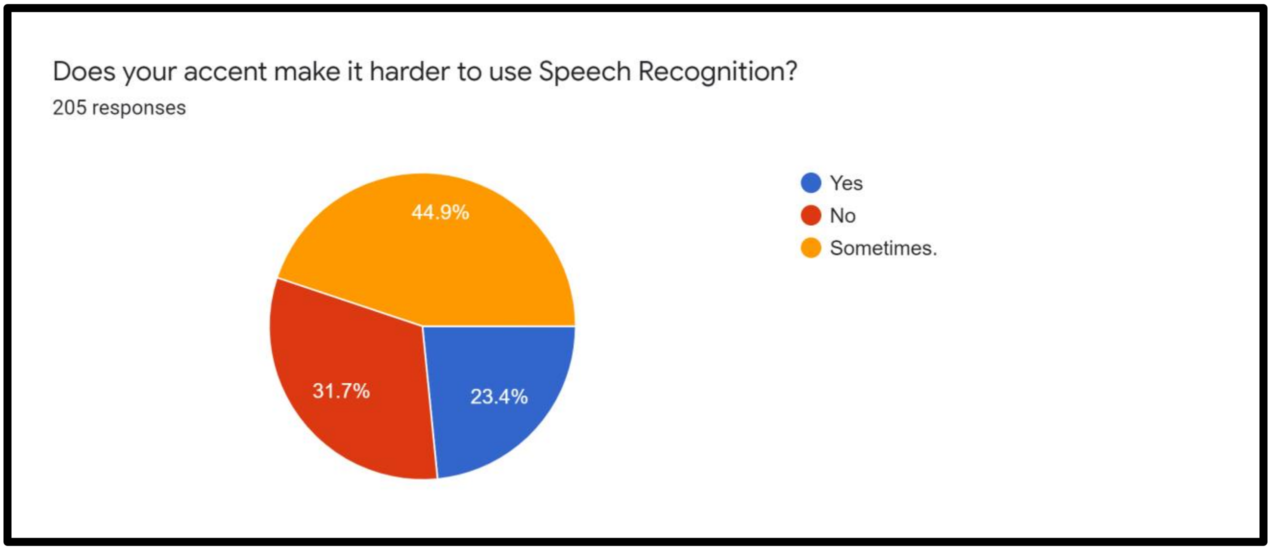


Figure 2.2.2. Accent and ASR

2.2.3. GENDER BIAS

Even though not intentional, speech recognition has serious race and gender bias which is problematic. Research by the North American Chapter of the Association for Computational Linguistics (NAACL) shows that Google’s speech recognition is 13% more accurate for men than it is for women. This is because the software used is programmed using the data collected from majorly American English-speaking males which makes it less responsive towards females, minors, and individuals of other races.

2.3. BACKGROUND NOISES

As we know acoustic and background conditions print a worse impact on speech recognition. Different acoustic conditions between training and evaluation data results in degradation of performance of speech recognition. The conditions which affect the quality of ASR mainly are:

2.3.1. ENVIRONMENTAL CONDITION

Disturbances such as loud noises, echoes and reverberation disrupt ASR Systems which makes them difficult to use outdoors and in urban areas.

2.3.2. AUDIO DEVICES

Devices used for speech recognition-devices used such as Microphone (far-field, close talking, directed, undirected), medium (noise cancellation disabled device). This can be rectified by the usage of a specific microphone or headset, which makes it undesirable.

2.3.3. PROPERTIES OF TRANSMITTING MEDIUM

Properties of audio transmitting medium (channel used, voice over, crosstalk).

2.3.4. AMBIENT EFFECTS

The adjustment of sound articulated in a noisy background (named as Lombard impact). This impact is intensely subject to the speaker, the specific circumstance and the degree of commotion making it hard to evaluate and modelized. All in all, foundation noise (mismatch among preparing and testing design) is the serious issue to manage in the negative impacts of background noises in ASR

CHAPTER 3: CONCLUSION

3.1. ACCURACY LEVELS

Word error rate has many faults and is mainly used as a marketing strategy. Its accuracy rate is a general way of showing how well the ASR performs but does not give us the technical aspects of its working. WER takes in only errors and does not factor in the variables causing the error. Moreover, many words are not given importance and all words are considered equal. To improve ASR, we need to include more factors other than WER and give a transparent view to the users of its functioning. Cross talk can also reduce the WER as the ASR generally omits the words which it cannot detect according to its data input. Speech data needs to be technical by adding audio with human-labelled transcripts. WER should be minimized to 3-5% as majorly its around 30% which in turn causes a lot of error.

To improve the search issue, the contextual analysis of the system needs to be improved so that the words which make better sense for the sentence that is being said by the user are added for efficient search and communication. In case of jargons more technical terms that are used in various paths of life are required to be added to the software for better efficiency.

ASR System analyse, recognize and compare phonemes. Not only is the process inefficient and time consuming but also the use of new words, slangs and jargon can significantly deteriorate the system’s performance.

3.2. ACCENT AND DISABILITIES

As ASR software is implemented in fields such as medicine and data input, improvement in accuracy levels is a must for automatic speech recognition. The software must also be made available for use among a wide range of individuals. This can be achieved by training the software further using speech models and audio inputs. However due to the unavailability of a variety of audio inputs the system is unable to grow further. This particular problem can be overcome by using audio input from media such as social media and movies, in speech models. This can enable the software to adopt accents, jargon, and learn to recognize unclear speech and disclose the margin between genders.

3.3. BACKGROUND NOISES

ASR performance quickly deteriorates in noisy conditions. Multiple methods have been proposed to cope with this problem such as speech signal acquisition, noise cancelling microphones and use of microphone arrays. The impact can be additionally diminished by the utilization of versatile or dynamic commotion dropping methods, division and discourse, non-discourse detection.

Albeit vigorous strategies must be planned since a few Noisy discourse acknowledgment mistakes begin from some unacceptable assurance of expression limits. Aside from these reference design displaying and acknowledgment calculations, distance estimates will likewise accumulate in diminishing the impact of foundation clamour on programmed discourse acknowledgment. Every one of these techniques joined can significantly improve the framework's acknowledgment precision.

CHAPTER 4: RECOMMENDATIONS

In the above discussion, we have seen the problems faced by people while using ASR and how frequently they use it. Along with the enhancement techniques in the ASR systems, we should also make people aware of what they can do from their side to get better outputs from the speech recognition systems.

- People should purchase better quality speakers or microphones so that the ASR system can interpret the input properly.

- Background noises can be subtracted by using noise cancellation devices so that the sounds that are relatively far away from the microphone are filtered out.

- The device’s microphone should be able to capture higher frequencies and detect the wavelength accordingly. This will help to distinguish between ambient noises in the background and the vocal inputs.

- The ASR system should be improved so that it can detect stutters automatically and can remember the speaker’s pace and fluency. This way people with speech disabilities like stutters can also use ASR.

- Instead of directly using speech recognition systems like Siri, people can use the ‘enable dictation’ option to send information and voice samples to the ASR database anonymously which can be used to get data input from a wide range of people. This technique helps the ASR system to learn new words or jargons from raw data that is fed into it.

CHAPTER 5: REFERENCES

[1] A. Aquino, J. L. Tsang, C. R. Lucas and F. de Leon, "G2P and ASR techniques for low-resource phonetic transcription of Tagalog, Cebuano, and Hiligaynon," 2019 International Symposium on Multimedia and Communication Technology (ISMAC), 2019, pp. 1-5, doi: 10.1109/ISMAC.2019.8836168.

[2] B. Worthy. “Word Error Rate Mechanism, ASR Transcription and Challenges in Accuracy Measurement.” gmrtranscription.com. https://www.gmrtranscription.com/blog/word-error-rate-mechanism-asr-transcription-and-challenges-in-accuracy-measurement

[3] H. Chen. “Does Word Error Rate Matter?” smartaction.ai.

https://www.smartaction.ai/blog/does-word-error-rate-matter/

[4] M. B. Mustafa, S. S. Salim, N. Mohamed, B. Al-Qatab, C. E. Siong. “Severity-based adaptation with limited data for ASR to aid dysarthric speakers. “PloS *one* vol.9,1 e86285. Jan. 2014, doi: 10.1371/journal.pone.0086285

[5] N. Jamal, S. Shanta, F. Mahmud, and M. Sha'abani, “Automatic speech recognition (ASR) based approach for speech therapy of aphasic patients: A review,” in *Advances in Electrical and Electronic Engineering: from Theory to Applications*, 2017, vol. 1883, no. 1. doi:10.1063/1.5002046.

[6] Ellen Eide, “Distinctive Features for Use in an Automatic Speech Recognition System,” in Proc. EuroSpeech 2001, Scandinavia, Aalborg; Denmark, 9 2001, ISCA.

[7] R. Tatman, “Google’s Speech Recognition has a Gender Bias,” makingnoiseandhearingthings.com. https://makingnoiseandhearingthings.com/2016/07/12/googles-speech-recognition-has-a-gender-bias.

[8] J. P. Bajorek,” Voice Recognition Still Has Significant Race and Gender Biases,” Harvard Business review.

https://hbr.org/2019/05/voice-recognition-still-has-significant-race-and-gender-biases.