An Efficient Method for Improving Automatic Speech Recognition

NUSRAT JAHAN[1], MD. ASHIKUR RAHMAN KHAN[1, *], ZAYED US SALEHIN[1],
NISHU NATH[1]
[1]Department of Information and Communication Engineering
Noakhali Science and Technology University
Noakhali-3814
BANGLADESH

Abstract: - Automatic speech recognition translates spoken words into the text; It is still a challenging task due to the high viability in speech signals. Several decoding algorithms and recognition systems have been developed, aimed at various recognition tasks. The design of the speech recognition system requires careful attention to the challenges or issue such as various types of speech classes, speech representation, feature extraction techniques, database and performance evaluation. This paper presents a study of basic approaches to speech recognition and also presents an error analysis of existing speech recognition system to provide a better system.

.*Keywords*: Hidden Markov model, Acoustic model, language model, Feature Extraction, Google Web Speech API, Voice Notepad.

Received: June 15, 2021. Revised: April 16, 2022. Accepted: May 17, 2022. Published: June 9, 2022.

1 Introduction

Speakers may have different accents, dialects or pronunciations, and speak in different styles, at different rates, and in different emotional states. . For more than half a century, research has been conducted in the field of automatic speech recognition (ASR), which constitutes an important part in the fulfilment of this vision. Despite the considerable amount of research resources invested in this task, many questions remain to be answered. This is because the problem is very complex and requires solutions from several disciplines. Early attempts at ASR were based on template matching techniques. HMMs are suitable for acoustic modelling in a context of concatenated acoustic-phonetic units. Although there are deficiencies associated with these devices for acoustic modelling, they have proven effective for the processing of continuous speech. The trend in ASR has been toward increasingly complex models, to improve recognition accuracy and involve larger vocabularies. This paper focuses on error analysis of speech recognition systems.

The trend in ASR has been toward increasingly complex models, to improve recognition accuracy and involve larger vocabularies. To handle these more complex recognition tasks, several advanced decoding strategies are required. The current challenges of speech recognition are caused by two major factors- reach and loud environments. The current challenges of speech recognitions are diverse. Speech recognition software isn't always able to interpret spoken words correctly. This is due to computers not being on par with humans in understanding the contextual relation of words and sentences, causing misinterpretations of what the speaker meant to say or achieve. People usually assume that computerizing a process would speed it up. Unfortunately, this is not always the case when it comes to voice recognition systems. In many cases using a voice, the app takes up more time than going with a traditional text-based version. While systems are getting better there’s still a big difference in their ability to understand American or Scottish English for example. Even a simple cold can be a reason for voice commands not to work as well as usual. keeping user data safe can easily become a conflict of interests**.** Therefore, a great challenge of voice recognition lies in making data input available for AI, but still, acknowledge the need for data privacy and security.

Some things have to take into consideration. These are

1. Variability in speech
2. Vocal Range (pitch and format Frequency)
3. Age, Gender of the speaker
4. Voice Quality
5. Emotional State
6. Speech Style

The main objective of this research is to study the accuracy, readability, accessibility of the speech recognition system. Also works with the problems or faults of the speech recognition system. In the speech, there is a lot to research. The specific objectives are:

1. To study the speech recognition system
2. Improve the accuracy of the speech recognition system
3. Error analysis
4. Make a comparison with the existed system
5. Use Speech recognition system in Education

2 Literature Survey

In here different related works are described in short. The followings are discussed, different types of existing research work.

A new acoustic model called Time-Inhomogeneous Hidden Bernoulli Model (TI-HBM) is introduced as an alternative to the Hidden Markov Model (HMM) in automatic speech recognition[1].

A significant new speech corpus of British English has been recorded at Cambridge University. Derived from the Wall Street Journal text corpus, WSJCAMO constitutes one of the largest corpora of spoken British English currently in existence. It has been specifically designed for the construction and evaluation of speaker-independent speech recognition systems. The database consists of 140 speakers each speaking about 110 utterances[2].

Deep recurrent neural networks investigate which combine the multiple levels of representation that have proved so effective in deep networks with the flexible use of long-range context that empowers RNN[3]. Using convolutional neural networks (CNNs)further error rate reduction can be obtained. Organization of the input Data to the CNN, CNN introduces a special network structure, which consists of alternating so-called convolution and pooling layers[4].A novel context-dependent (CD) model for large-vocabulary speech recognition (LVSR) that leverages recent advances in using deep belief networks for phone recognition[5].A vector Taylor series approach for environment-independent speech recognition. In this paper, they introduce a new analytical approach to environment compensation for speech recognition. In this work also introduce the use of a Vector Taylor series (VTS) expansion to characterize efficiently and accurately the effects on speech statistics of unknown additive noise and unknown linear filtering in a transmission channel[6]. Word speech recognition algorithm based on hidden Markov models (HMM's) describes the modifications made to a connection which allow it to recognize words from a predefined vocabulary list spoken in an unconstrained fashion[7]. They propose to apply CNN to speech recognition within the framework of hybrid NN-HMM model. They propose to use local filtering and max-pooling in the frequency domain to normalize speaker variance to achieve higher multi-speaker speech recognition performance. A pair of local filtering layer and the max-pooling layer is added at the lowest end of the neural network (NN) to normalize spectral variations of speech signals. The proposed CNN architecture is evaluated in a speaker-independent speech recognition task using the standard TIMIT data sets. Experimental results show that the proposed CNN method can achieve over 10% relative error reduction in the core TIMIT test sets when comparing with a regular NN using the same number of hidden layers and weights[8].

A statistical modelling procedure that was developed to account for the fact that, in a forensic voice comparison analysis conducted for a particular case, there was a long time interval between when the questioned- and known-speaker recordings were made (six years), but in the sample of the relevant population used for training and testing the forensic voice comparison system, there was a short interval (hours to days) between when each of multiple recordings of each speaker was made. The present paper also includes the results of empirical validation of the procedure [9]. Bimodal recognition Speech recognition and speaker recognition by machine are crucial ingredients for many important applications such as natural and flexible human-machine interfaces. Most developments in speech-based automatic recognition have relied on acoustic speech as the sole input signal, disregarding its visual counterpart [10]. Many state-of-the-art Large Vocabulary Continuous Speech Recognition (LVCSR) systems are hybrids of neural networks and Hidden Markov Models (HMMs). Recently, more direct end-to-end methods have been investigated, in which neural architectures were trained to model sequences of characters [11]. A method of compensating for nonlinear distortions in speech representation describes caused by noise. The paper shows how the proposed method can be applied to robust speech recognition and it is compared with other compensation techniques[12].In this paper, they extend the interpretation of distortion measures, based upon the observation that measurements of speech spectral envelopes (as normally obtained from standard analysis procedures such as LPC or filter banks) are prone to statistical variations due to window position fluctuations, excitation interference, measurement noise, etc., and may not accurately characterize the true speech spectrum because of analysis model constraints. They have found that a bandpass "liftering" process reduces the variability of the statistical components of LPC-based spectral measurements and hence it is desirable to use such a liftering process in a speech recognizer. Using the liftering process, they have been able to achieve an average digit error rate of 1 per cent in a speaker-independent isolated digit test. This error rate is about one-half that obtained without the liftering process [13].A new technique for training deep neural networks (DNNs) as data-driven feature front-ends for large vocabulary continuous speech recognition (LVCSR) in low resource settings. In their experiments, the proposed features provide an absolute improvement of 16% in a low-resource LVCSR setting with only one hour of in-domain training data. While close to three-fourths of these gains come from DNN-based features, the remaining are from semi-supervised training [14]. The application of deep neural network (DNN) - hidden Markov model (HMM) hybrid acoustic models for far-field speech recognition of meetings recorded using microphone arrays. They investigate the application of deep neural network (DNN)-hidden Markov model (HMM) hybrid acoustic models for far-field speech recognition of meetings recorded using microphone arrays. They show that hybrid models achieve significantly better accuracy than conventional systems based on Gaussian mixture models (GMMs) [15]. A novel recurrent neural network (RNN) model for voice activity detection has presented. Their multi-layer RNN model, in which nodes compute quadratic polynomials, outperforms a much larger baseline system composed of Gaussian mixture models (GMMs) and a hand-tuned state machine (SM) for temporal smoothing [16].

3 Methodology

Hidden Markov model, Acoustic model, Language model, feature extraction that is used in the speech recognition system. Human speech production is an important matter in the speech recognition system.

3.1 Human Speech Production

Speech, being a tool of communication, is also a symbol of identity and authorization. Speech is produced through different parts of the mouth that creates air pressure. The changes can then be sampled periodically and recorded in a digital waveform. The waveform carries all the information of the spoken word. The physical shape of a human vocal tract is different from a person by person. Hence, each human speaks differently. The environment where human speaks, the dialect of the language, differences in the vocal tract length of males, female and children provide the speech variation and thus make it difficult to understand speech signals. However, there are still some features in the human speech which can be mathematically modelled and used for predicting words from it but it demands a tremendous amount of time and effort. Most of the speech recognition systems today use statistical models. A large set of model training data is used to calculate the features. The statistical model requires acoustic modelling. Acoustic modelling is represented by the Hidden Markov Model.

3.2 Used Model

Hidden Markov Model (HMM) is a statistical Markov model in which the system being modelled is assumed to be a Markov process with unobserved (i.e. *hidden*) states. In machine learning, pattern recognition and in image processing, feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations.MFCC feature extraction: The MFCC feature extraction technique includes windowing the signal, applying the DFT, taking the log of the magnitude, and then warping the frequencies on a Mel scale, followed by applying the inverse DCT.The acoustic model is used in automatic speech recognition to represent the relationship between an audio signal and the phonemes or other linguistic units that make up speech. It has been established that acoustic models of speech recognition capture the characteristics of the basic recognition units phoneme is the most favourable unit. A statistical language model is a probability distribution over sequences of words One of the major objectives of the language model is to convey or transmit the behaviour of the language.

3.3 Proposed Speech Recognition System

Here, the proposed method of recognition systems is given below. In this method, there have focused on the perspective and other aspects which affect the performance of automatic speech recognition.

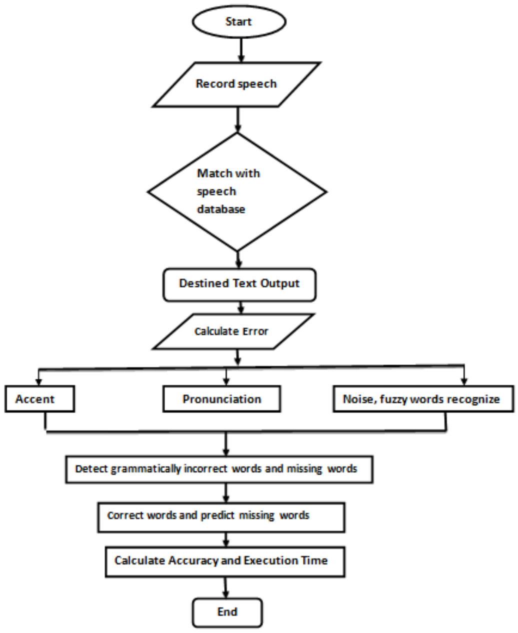

Start
Record speech
Match with speech database
Destined Text Output
Calculate Error
Accent
Pronunciation
Noise, fuzzy words recognize
Detect grammatically incorrect words and missing words
Correct words and predict missing words
Calculate Accuracy and Execution Time
End

Figure 1. The proposed method of the speech recognition system

4 Implementation, Result Analysis and Discussion

4.1 Implementation

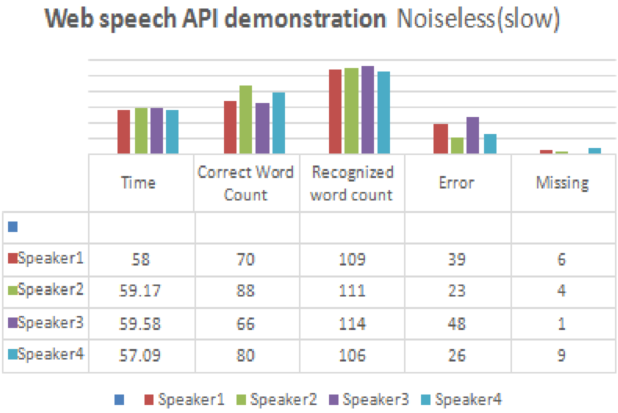
A comparative analysis by using different software of speech recognition has been done. First, here used Web speech API demonstration and Voice Notepad - Speech to Text with Google Speech Recognition. The experiment has been done in both Noise-less and noisy environment. Both environment, words are spoken by the speakers in two ways. First, words are spoken in slower precess. And then words are also spoken in a faster way. Error analysis of two speech recognition system has been made. According to noise and noiseless environment the correct words, recognized word, missing words and error words are counted. Four speakers are used for this analysis. There have used both the slow and fast of speaking words. Here, errors have been done more in a fast-noisy environment. Sometimes, the speech recognition system cannot recognize words even in the noiseless slow environment. But, the accuracy of understanding the words are more efficient in the noiseless slow environment. Here, a passage contains 115 words are used. Some speaker needs more time and some speaker needs less time to finish reading the passage.

Comparatively, both the Web speech API demonstration and Voice Notepad - Speech to Text with Google Speech Recognition have given the almost same accuracy.

There have done it in four ways. These are

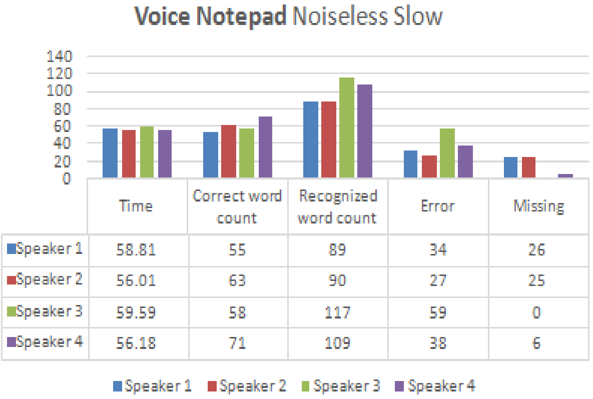
1. Noiseless slow
2. Noiseless Fast
3. Noisy slow
4. Noisy Fast

From figure (2-9)there have shown the comparison of correct word count, recognize word count, Error words and missing words according to time.



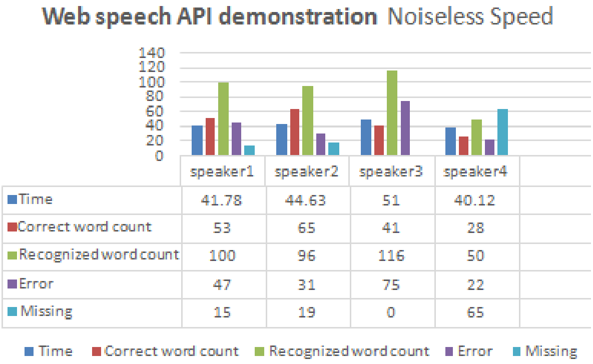
	Time	Correct Word Count	Recognized word count	Error	Missing
🟦					
🟥Speaker1	58	70	109	39	6
🟩Speaker2	59.17	88	111	23	4
🟪Speaker3	59.58	66	114	48	1
🟦Speaker4	57.09	80	106	26	9

Figure 2. Web Speech API demonstrationNoiseless Slow



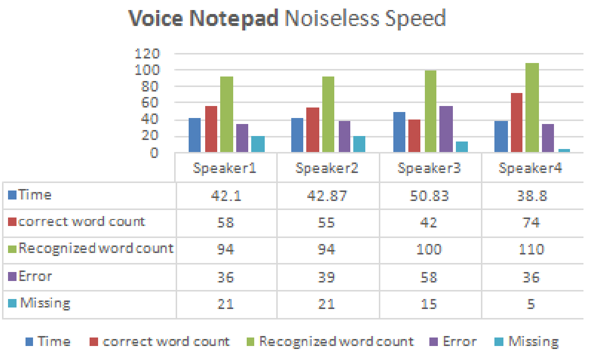
	Time	Correct word count	Recognized word count	Error	Missing
🟦Speaker 1	58.81	55	89	34	26
🟥Speaker 2	56.01	63	90	27	25
🟩Speaker 3	59.59	58	117	59	0
🟪Speaker 4	56.18	71	109	38	6

Figure 3.Voice Notepad, Noiseless Slow.



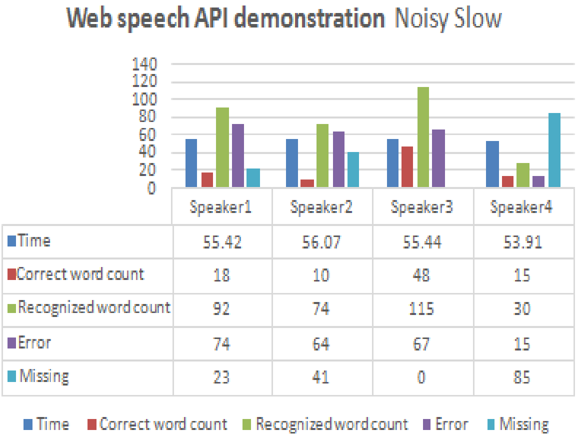
	speaker1	speaker2	speaker3	speaker4
🟦Time	41.78	44.63	51	40.12
🟥Correct word count	53	65	41	28
🟩Recognized word count	100	96	116	50
🟪Error	47	31	75	22
🟦Missing	15	19	0	65

Fig 4: Web Speech API demonstrationNoiseless Speed



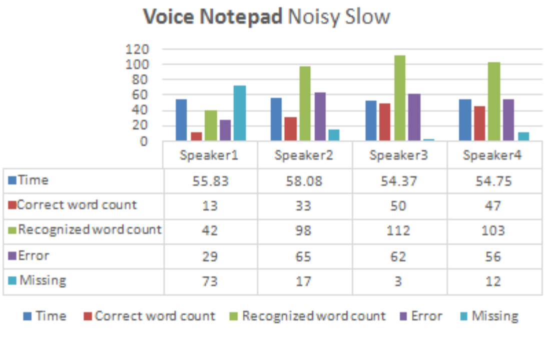
	Speaker1	Speaker2	Speaker3	Speaker4
Time	42.1	42.87	50.83	38.8
correct word count	58	55	42	74
Recognized word count	94	94	100	110
Error	36	39	58	36
Missing	21	21	15	5

Fig 5: Voice Notepad Noiseless Speed

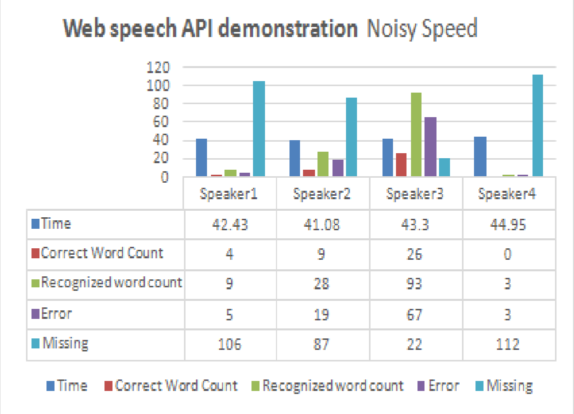


	Speaker1	Speaker2	Speaker3	Speaker4
Time	55.42	56.07	55.44	53.91
Correct word count	18	10	48	15
Recognized word count	92	74	115	30
Error	74	64	67	15
Missing	23	41	0	85

Fig 6: Web Speech API demonstration Noisy Slow

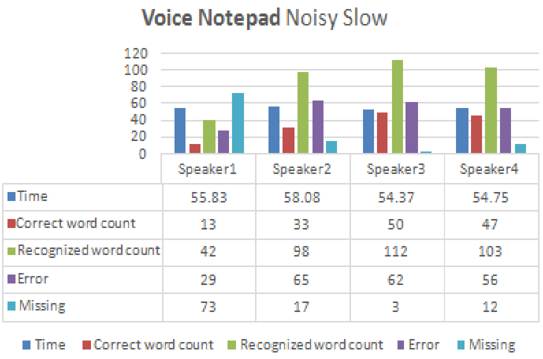


	Speaker1	Speaker2	Speaker3	Speaker4
Time	55.83	58.08	54.37	54.75
Correct word count	13	33	50	47
Recognized word count	42	98	112	103
Error	29	65	62	56
Missing	73	17	3	12



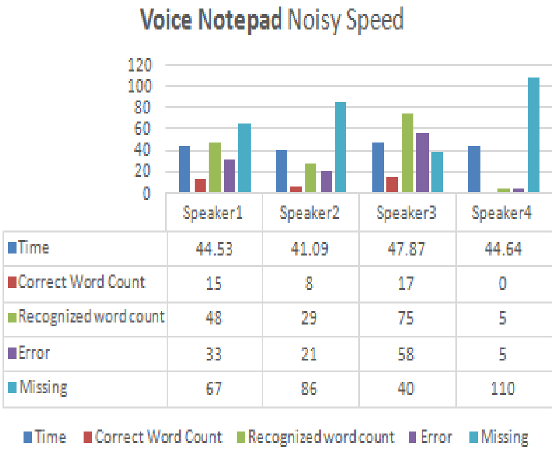
	Speaker1	Speaker2	Speaker3	Speaker4
Time	42.43	41.08	43.3	44.95
Correct Word Count	4	9	26	0
Recognized word count	9	28	93	3
Error	5	19	67	3
Missing	106	87	22	112

Fig 7: Web Speech API demonstrationNoisy Speed



	Speaker1	Speaker2	Speaker3	Speaker4
Time	55.83	58.08	54.37	54.75
Correct word count	13	33	50	47
Recognized word count	42	98	112	103
Error	29	65	62	56
Missing	73	17	3	12

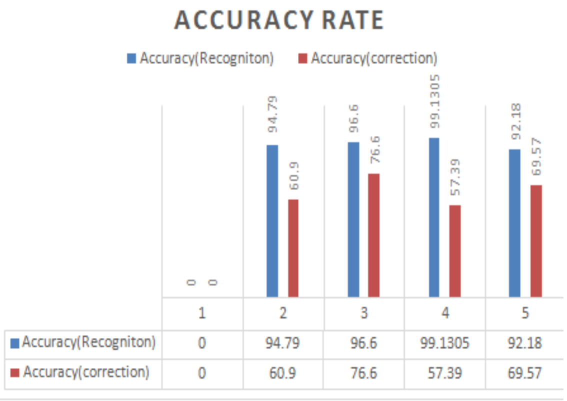
Fig 8: Voice Notepad Noisy Slow



	Speaker1	Speaker2	Speaker3	Speaker4
Time	44.53	41.09	47.87	44.64
Correct Word Count	15	8	17	0
Recognized word count	48	29	75	5
Error	33	21	58	5
Missing	67	86	40	110

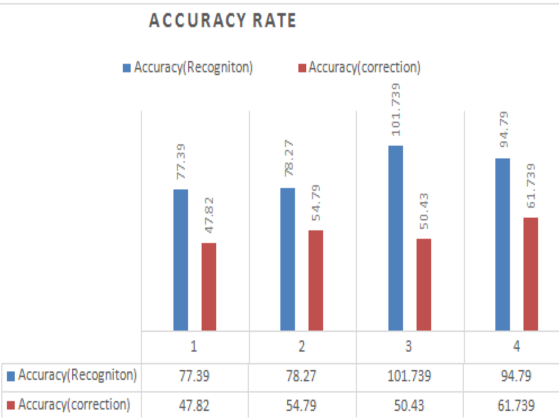
Fig 9: Voice Notepad Noisy Speed

Then, Below from figure (11-17), there have shown the accuracy rate of 4 speakers in these 4 environments in a statistical manner. There stepped up by word recognition and correct word recognition. Accuracy rate varies with from person to person. There have also seen that word recognition accuracy is better than correct word recognition. Sometimes there produced different words but not the same word.



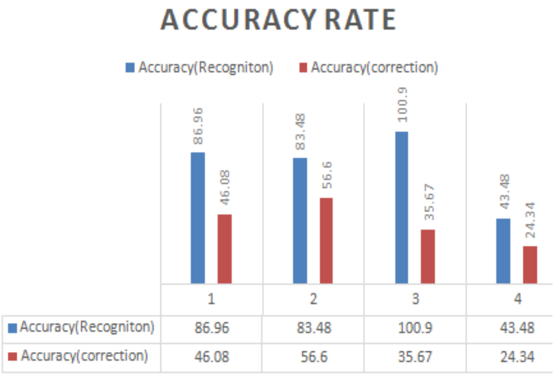
	1	2	3	4	5
Accuracy(Recogniton)	0	94.79	96.6	99.1305	92.18
Accuracy(correction)	0	60.9	76.6	57.39	69.57

Fig 10: Web Speech API (Noiseless slow)



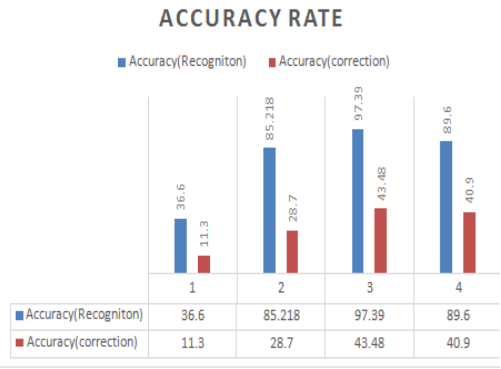
	1	2	3	4
Accuracy(Recogniton)	77.39	78.27	101.739	94.79
Accuracy(correction)	47.82	54.79	50.43	61.739

Fig 11: Notepad slow Noiseless



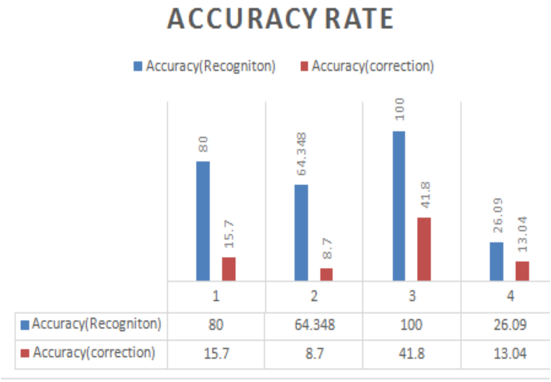
	1	2	3	4
Accuracy(Recogniton)	86.96	83.48	100.9	43.48
Accuracy(correction)	46.08	56.6	35.67	24.34

Fig 12: Web Speech API (speed Noiseless)



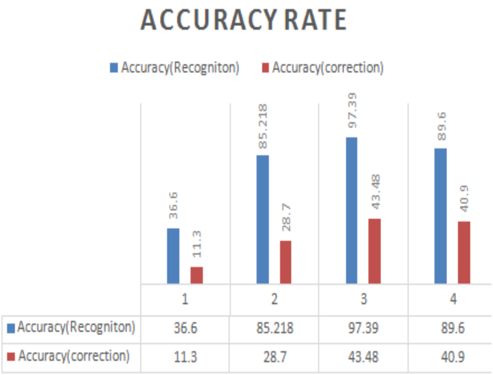
	1	2	3	4
Accuracy(Recogniton)	36.6	85.218	97.39	89.6
Accuracy(correction)	11.3	28.7	43.48	40.9

Fig 13: Notepad (speed Noiseless)



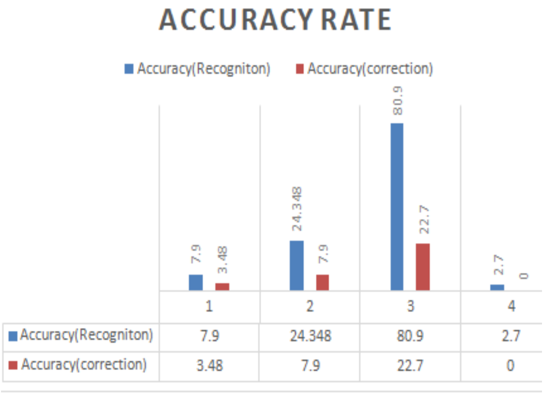
	1	2	3	4
Accuracy(Recogniton)	80	64.348	100	26.09
Accuracy(correction)	15.7	8.7	41.8	13.04

Fig 14: Web Speech API (slow noisy)



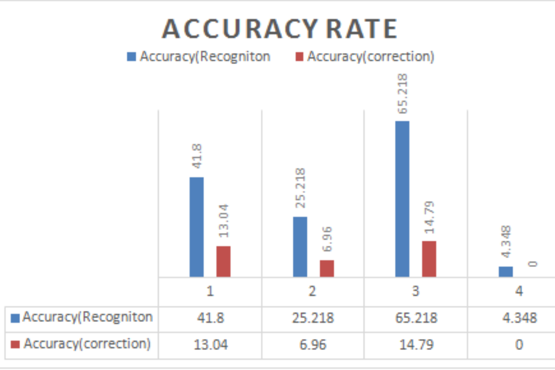
	1	2	3	4
Accuracy(Recogniton)	36.6	85.218	97.39	89.6
Accuracy(correction)	11.3	28.7	43.48	40.9

Fig 15: Notepad (slow noisy)



	1	2	3	4
Accuracy(Recogniton)	7.9	24.348	80.9	2.7
Accuracy(correction)	3.48	7.9	22.7	0

Fig 16: API speed noisy



	1	2	3	4
Accuracy(Recogniton	41.8	25.218	65.218	4.348
Accuracy(correction)	13.04	6.96	14.79	0

Fig 17: Notepad speed Noisy

4.2 Result Analysis and Discussion

This can be done by using another platform from also. There have shown that in the existing system how it responds to the environment and also the recognition rates are varied from human to human. In this research work, error analysis plays a vital part. Due to different types of obstacles during a speech to text conversion outputs have been analysed. Some difficulties with ASR is Human comprehension of speech, spoken language is not equal to written language, Noise, channel Variability, realization, dialects. For example, in the analysis here, 4 speakers have 4 different types of accent, pronunciation. For this reason, although there have used the same paragraph, recognizing the word numbers are not same. Same words, by speaking one human-computer can recognize but by speaking another human, it cannot recognize. Here, fuzzy and complicated words are not recognized by the computer. In some situation, it produced error words and sometimes words are missing. Before the solution looks through the problems are essential. There have tried to show the problems of the existing speech recognition system. For reducing these problems a system has been proposed. In this paper, existing system problems have analysed so that they can show the problems in the speech recognition system. There should be a method which will be cheap and can recognize almost all types of words. That can help us to use the speech recognition system more properly in our daily life. Just as there used a touch screen mobile phone or Microsoft office in everyday life.

5 Conclusions

The most important step in all of this research is the increasing inaccessibility. There is no use increasing the readability and accuracy of Speech-to-Text Transcripts if they will not be used. Systems should be easy to implement and demonstrated throughout our everyday lives.

Reference

[1] Jahanshah, Kabudian,M.Mehdi, Homayounpour, S.Mohammad Ahadi "Time-inhomogeneous hidden Bernoulli model: An alternative to hidden Markov model for automatic speech recognition", *2008 International Conference on Acoustics, Speech and Signal Processing, IEEE*,(ISSN) 2379-190X,2008.

[2] T. Robinson, J. Fransen, D. Pye, J. Foote, S. Renals "WSJCAMO: a British English speech corpus for large vocabulary continuous speech recognition" *1995 International Conference on Acoustics, Speech, and Signal Processing, IEEE*, Vol. 1, pp. 81-84,1995.

[3] Alex Graves, Abdel-rahman Mohamed, Geoffrey Hinton "Speech recognition with deep recurrent neural networks", *2013 IEEE international conference on acoustics, speech and signal processing,***(**ISSN)1520-6149, pp. 6645-6649,2013.

[4] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn and Dong Yu, "Convolutional neural networks for speech recognition", *IEEE/ACM Transactions on audio, speech, and language processing*, Vol. 22, No.10, pp.1533-1545, Oct. 2014.

[5] George E. Dahl, Dong Yu, Li Deng, Alex Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition" *IEEE Transactions on Audio, Speech, and Language Processing*, Vol:20, No.1, pp.30-42,2012.

[6] P.J. Moreno, B. Raj, "A vector Taylor series approach for environment-independent speech recognition", *1996IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, Vol. 2, pp. 733-736,1996

[7] J.G. Wilpon, L.R. Rabiner, C.-H. Lee, E.R. Goldman, "Automatic recognition of keywords in unconstrained speech using hidden Markov models", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol.38, No.11, pp.1870-1878,1990.

[8] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Gerald Penn, "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition", *2012 IEEE international conference on Acoustics, speech and signal processing (ICASSP)*, pp. 4277-4280,2012.

[9] Geoffrey Stewart Morrison, Finnian Kelly, "A statistical procedure to adjust for time-interval mismatch in forensic voice comparison", *Speech Communication*, Vol.112, pp.15-21,2019.

[10] C.C. Chibelushi, F. Deravi, J.S.D. Mason, "A review of speech-based bimodal recognition",*IEEE transactions on multimedia*, Vol:4,No.1, pp.23-37,2002.

[11]. Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philémon Brakel, Yoshua Bengio, "End-to-end attention-based large vocabulary speech recognition", *2016IEEE international conference on acoustics, speech and signal processing (ICASSP),(*ISSN)2379-190X, pp. 4945-4949, 2016.

[12] De A. de la Torre, A.M. Peinado, J.C. Segura, J.L. Perez-Cordoba, M.C. Benitez, A.J. Rubio, "Histogram equalization of speech representation for robust speech recognition", *IEEE Transactions on Speech and Audio Processing*, vol.13, No.3, pp.355-366,2005.

[13] Biing-Hwang Juang, L. Rabiner, J. Wilpon, "On the use of bandpass liftering in speech recognition", *IEEE Transactions on acoustics,*

speech, and signal processing, Vol.35, No.7, pp.947-954,1987.

[14] Samuel Thomas, Michael L. Seltzer, Kenneth Church, Hynek Hermansky, "Deep neural network features and semi-supervised training for low resource speech recognition",*2013 IEEE international conference on acoustics, speech and signal processing*,(ISSN) 2379-190X, pp. 6704-6708,2013

[15] Pawel Swietojanski, Arnab Ghoshal, Steve Renals, "Hybrid acoustic models for distant and multichannel large vocabulary speech recognition", *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 285-290,2013

[16] Thad Hughes, Keir Mierle, "Recurrent neural networks for voice activity detection", *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*,(ISSN) 2379-190X, pp.7378-7382,2013.