International Journal of Early Childhood Special Education (INT-JECSE)
DOI:10.48047/INTJECSE/V15I4.93 ISSN: 1308-5581 Vol 15, Issue 04 2023

SPEECH TO TEXT RECOGNTION

Mr. M.SUDHAKAR
ASSISTANT PROFESSOR
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SREYAS INSTITUTE OF ENGINEERING AND TECHNOLOGY
Mail id: sudhakar.m@sreyas.ac.in

RAM SAI SANTOSH MANI TEJA
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SREYAS INSTITUTE OF ENGINEERING AND TECHNOLOGY
Mail id: siddasantosh95@gmail.com

ERRA DHARSHITH
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SREYAS INSTITUTE OF ENGINEERING AND TECHNOLOGY
Mail id:dharshithsunny@gmail.com

KOTTINTI AKSHAYA
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SREYAS INSTITUTE OF ENGINEERING AND TECHNOLOGY
Mail id:kothintiakshaya@gmail.com

VAKITI SONY
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SREYAS INSTITUTE OF ENGINEERING AND TECHNOLOGY
Mail id:sonyvakiti386@gmail.com

ABSTRACT

Human voice synthesis technology is one from the fast-growing engineering technologies. It has a number of applications in different areas and provides potential benefits. Nearly 20% people of the world are suffering from various disabilities, many of them are blind or unable to use their hands effectively. The speech recognition systems in those particular cases provide a significant help to them, so that they can share information with people by operating computer through voice input. This project is designed and developed keeping that factor into mind, and a little effort is made to achieve this aim. Our project is capable to recognize the human speech and convert the input audio into text.

INTRODUCTION

In the fast-paced digital era, technological innovations continue to reshape the way we interact with information. One such groundbreaking advancement that has transformed the landscape of communication and accessibility is Speech-to-Text Recognition technology. This transformative technology, often abbreviated as STT or ASR (Automatic Speech Recognition), holds the promise of breaking down language barriers, enhancing accessibility for individuals with disabilities, revolutionizing customer service, and significantly improving productivity across various sectors. This introduction delves into the evolution, significance, and the far-reaching impact of Speech-to-Text Recognition technology on society, education, business, and beyond.

The concept of converting spoken language into written text has fascinated researchers and inventors for decades. Early attempts at speech recognition date back to the mid-20th century when scientists explored rudimentary methods to decipher spoken words. However, it is in recent years, thanks to the exponential growth in computational power and machine learning algorithms, that Speech-to-Text Recognition technology has made remarkable strides. Modern applications of this technology encompass a wide array of domains, from mobile devices and virtual assistants to transcription services and language translation tools.

One of the most significant contributions of Speech-to-Text Recognition technology is its ability to bridge communication gaps. For individuals with hearing impairments or speech disabilities, STT systems serve as invaluable tools, enabling them to participate in conversations, access information, and engage in educational and professional pursuits. This inclusivity empowers individuals who were previously marginalized to communicate effectively, fostering a more equitable society.

In the realm of education, Speech-to-Text Recognition technology has revolutionized the learning experience. Students with learning disabilities can now receive real-time transcription of lectures, making classroom content more accessible. Additionally, educators can create interactive and engaging learning environments by incorporating voice-activated technologies into educational software, thereby enhancing student participation and comprehension.

Businesses have also embraced Speech-to-Text Recognition technology to enhance customer service and streamline operations. Interactive voice response (IVR) systems powered by speech recognition facilitate seamless customer interactions, allowing users to navigate menus, make inquiries, and conduct transactions using spoken language. Moreover, in sectors such as healthcare, legal, and journalism, where accurate and timely transcription is paramount, STT technology has significantly reduced the time and effort required for documentation, leading to increased efficiency and productivity.

The globalized world demands effective communication across languages and cultures. Speech-to-Text Recognition technology plays a vital role in breaking down language barriers by providing real-time translation services. Whether in international business negotiations, diplomacy, or travel, individuals can communicate effortlessly in their native languages, transcending linguistic boundaries and fostering cross-cultural understanding.

As Speech-to-Text Recognition technology continues to evolve, its future implications are vast and promising. However, challenges persist, including dialectal variations, background noise, and context understanding. Researchers and engineers are tirelessly working to overcome these hurdles, aiming to create more robust, accurate, and context-aware STT systems. Furthermore, ethical considerations, such as user privacy and data security, are paramount as the technology becomes more pervasive in our daily lives.

In conclusion, Speech-to-Text Recognition technology stands at the forefront of the technological revolution, transforming the way we communicate, learn, conduct business, and connect with one another. Its impact on accessibility, education, customer service, and global communication is unparalleled, fostering inclusivity and innovation across diverse sectors. As researchers and innovators continue to push the boundaries of this transformative technology, society can anticipate a future where seamless, natural language interaction becomes the norm, enriching lives and fostering a more connected and understanding world.

LITERATURE SURVEY

“Speech Recognition with Hidden Markov Model”.Speech Recognition means interpreting voice of the computer and performing given task or ability to match a voice against a available or given vocabulary. The actual task is to make the computer to understand spoken language. By “Understand” we intend to respond fittingly and change over the info discourse into another medium e.g.text. Speech recognition is therefore sometimes called as speech-to-text (STT). A speech recognition framework comprises of a mouthpiece, for the individual to talk into and speech recognition programming; a computer to take and interpret the speech; a good quality soundcard for input and/or output; a proper and good pronunciation. If we had a computer system which can do half as decent a job of recognizing human speech as other human beings can, and do it economically, speech will eventually replace cards, paper tape and even keyboards for communication with computers. The technique of automatic speech recognition has improved remarkably in the past decade. With the development in accuracy and scope, there has come, for the time being, a strong juncture on a class of statistical methods based on a structure called a hidden Markov model (HMM).

"Synthesis and evaluation of conversational characteristics in HMM-based speech synthesis," Spontaneous conversational speech has many characteristics that are currently not modelled well by HMM-based speech synthesis and in order to build synthetic voices that can give an impression of someone partaking in a conversation, we need to utilise data that exhibits more of the speech phenomena associated with conversations than the more generally used carefully read aloud sentences. In this paper we show that synthetic voices built with HMM-based speech synthesis techniques from conversational speech data, preserved segmental and prosodic characteristics of frequent conversational speech phenomena. An analysis of an evaluation investigating the perception of quality and speaking style of HMM-based voices confirms that speech with conversational characteristics are instrumental for listeners to perceive successful integration of conversational speech phenomena in synthetic speech. The achieved synthetic speech quality provides an encouraging start for the continued use of conversational speech in HMM-based speech synthesis.

"Voice recognition technology as a tool for behavioral research" Behavioral research often requires the acquisition and processing of large volumes of data. Most current techniques for recording behavior constrain the amount and type of data that can be measured. We developed and tested a system that uses voice recognition technology to collect data on the social interactions and singing patterns of cowbirds (Molothrus ater) living outdoors in a semi-natural environment. We spoke observation data into a wireless microphone that transmitted the data to a computer in the laboratory. After collection, the data were automatically checked for errors and then were entered into a database. Overall, the system performed at extremely high levels of accuracy. We tested the system under the challenging circumstances of field observation, and it performed above our expectations. If transmission difficulties are removed, voice recognition could be even more accurate. We recommend voice recognition as a powerful new tool for the variety of research fields in which measuring behavior is involved.

"Voice recognition based wireless home automation system" Home Automation industry is growing rapidly; this is fuelled by the need to provide supporting systems for the elderly and the disabled, especially those who live alone. Coupled with this, the world population is confirmed to be getting older. Home automation systems must comply with the household standards and convenience of usage. This paper details the overall design of a wireless home automation system (WHAS) which has been built and implemented. The automation centres on recognition of voice commands and uses low-power RF ZigBee wireless communication modules which are relatively cheap. The home automation system is intended to control all lights and electrical appliances in a home or office using voice commands. The system has been tested and verified. The verification tests included voice recognition response test, indoor ZigBee communication test, and the compression and decompression tests of DPCM (Differential Pulse Code Modulation) speech signals. The tests involved a mix of 35 male and female subjects with different English accents. 35 different voice commands were sent by each person. Thus the test involved sending a total of 1225 commands and 79.8% of these commands were recognised correctly.

"Enriching speech recognition with automatic detection of sentence boundaries and disfluencies" Survey: Effective human and automatic processing of speech requires recovery of more than just the words. It also involves recovering phenomena such as sentence boundaries, filler words, and disfluencies, referred to as structural metadata. We describe a metadata detection system that combines information from different types of textual knowledge sources with information from a prosodic classifier. We investigate maximum entropy and conditional random field models, as well as the predominant hidden Markov model (HMM) approach, and find that discriminative models generally outperform generative models. We report system performance on both broadcast news and conversational telephone speech tasks, illustrating significant performance differences across tasks and as a function of recognizer performance. The results represent the state of the art, as assessed in the NIST RT-04F evaluation.

PROPOSED SYSTEM

Speech recognition technology has made remarkable strides in recent years, transforming the way humans interact with machines. The ability to convert spoken language into written text, often referred to as Speech to Text (STT) recognition, has found applications in various fields, including transcription services, voice assistants, and accessibility tools. OpenCV, a popular computer vision library, has been instrumental in advancing speech recognition capabilities, providing robust tools for processing audio signals and converting them into textual information. This article explores the operation of Speech to Text recognition using OpenCV, shedding light on the underlying processes and demonstrating its significance in the realm of human-computer interaction.Speech to Text recognition involves the conversion of spoken language into written text. This process relies on sophisticated algorithms and machine learning models that analyze audio signals, identify speech patterns, and transform them into textual form. The goal is to bridge the gap between spoken words and written language, enabling seamless communication between humans and machines.

OpenCV, primarily known for its computer vision capabilities, plays a crucial role in audio signal processing, a fundamental aspect of speech recognition. OpenCV provides a range of tools for reading, processing, and analyzing audio data. By leveraging Open CV's functionality, developers can preprocess audio signals effectively, extracting features essential for speech recognition tasks. Speech recognition begins with capturing audio input. This input can come from various sources, such as microphones, audio files, or streaming services. OpenCV facilitates audio input by providing interfaces to access live audio streams or pre-recorded audio files. Raw audio signals often contain noise and irrelevant information that can hinder the accuracy of speech recognition. OpenCV offers functions for noise reduction, filtering, and feature extraction. Preprocessing steps enhance the quality of the audio signal, making it suitable for analysis.

OpenCV provides tools for extracting essential features from audio signals, such as Mel-frequency cepstral coefficients (MFCCs) and spectrograms. These features capture the spectral characteristics of the audio, representing them as numerical data that machine learning models can process effectively. Open CV integrates seamlessly with machine learning frameworks like scikit-learn and Tensor Flow. Developers can use these frameworks to train models, such as deep neural networks and recurrent neural networks, on the extracted audio features. These models learn to recognize patterns in the audio data, associating them with specific words or phrases.

Once the machine learning model processes the audio features, it generates text output corresponding to the recognized speech. OpenCV facilitates the conversion of numerical results back into human-readable text, providing the final output of the speech recognition process. Speech to Text recognition, enabled by OpenCV, enhances accessibility for individuals with disabilities. It allows those with speech impairments to communicate effectively by converting their spoken words into written text.

In various industries, such as journalism, legal, and healthcare, efficient transcription of spoken words into text is crucial. OpenCV-based Speech to Text recognition automates the transcription process, saving time and resources.

Voice assistants like Siri, Alexa, and Google Assistant rely on Speech to Text recognition to understand user commands and queries. OpenCV-powered speech recognition enhances the accuracy and responsiveness of these voice-driven interfaces, improving the overall user experience.Combining speech recognition with other modalities, such as image or video processing, enables the development of multimodal applications. Open CV's versatility allows developers to create innovative applications that integrate speech recognition seamlessly.

Speech to Text recognition, a transformative technology, has become an integral part of our daily lives, revolutionizing the way we interact with machines. OpenCV, with its powerful audio signal processing capabilities, serves as a cornerstone in the development of accurate and efficient speech recognition systems. By understanding the operation of Speech to Text recognition using OpenCV, developers can harness the potential of this technology to create inclusive applications, streamline transcription services, enhance voice assistants, and explore new horizons in multimodal human-computer interaction. As technology continues to advance, the synergy between speech recognition and OpenCV promises exciting possibilities, paving the way for more natural and intuitive interactions between humans and machines.

DOI:10.48047/INTJECSE/V15I4.93 ISSN: 1308-5581 Vol 15, Issue 04 2023

RESULTS

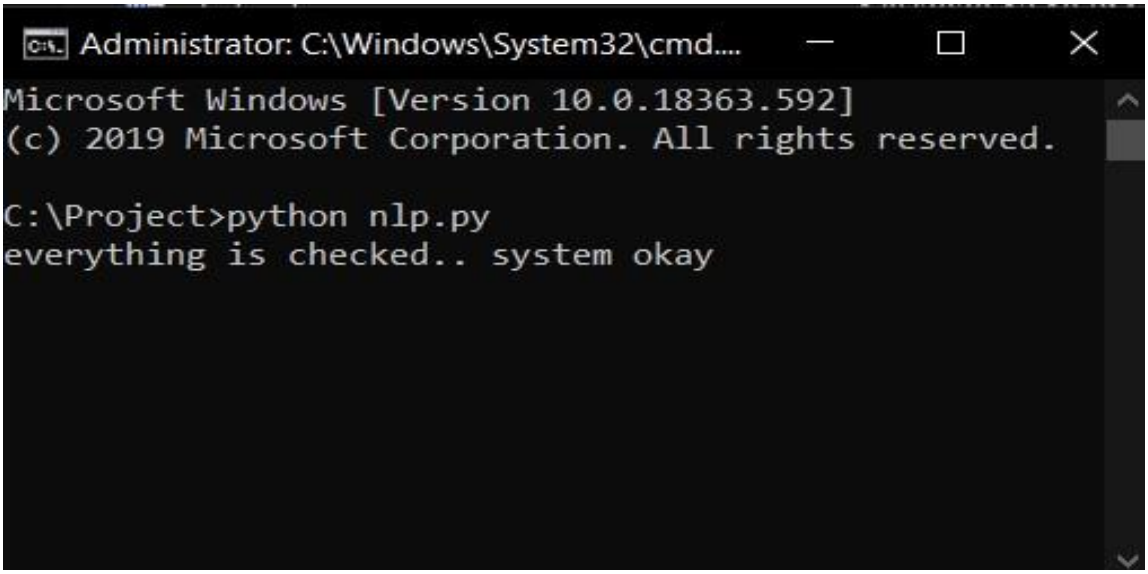

Fig: 1 Input Command

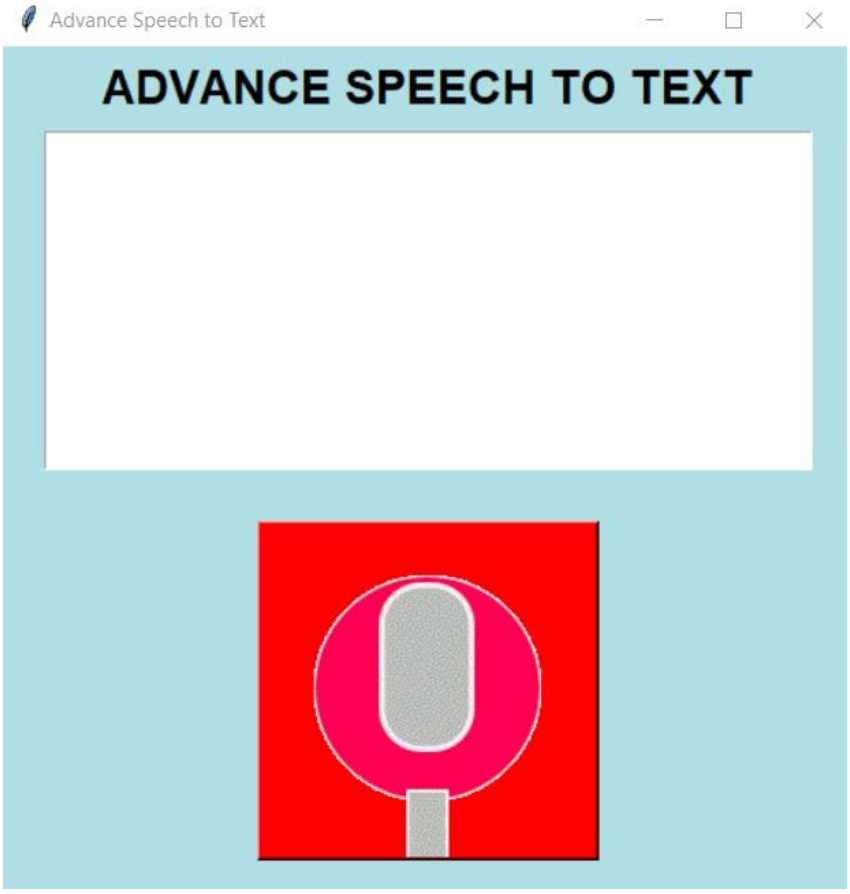

Fig: 2 Output Model

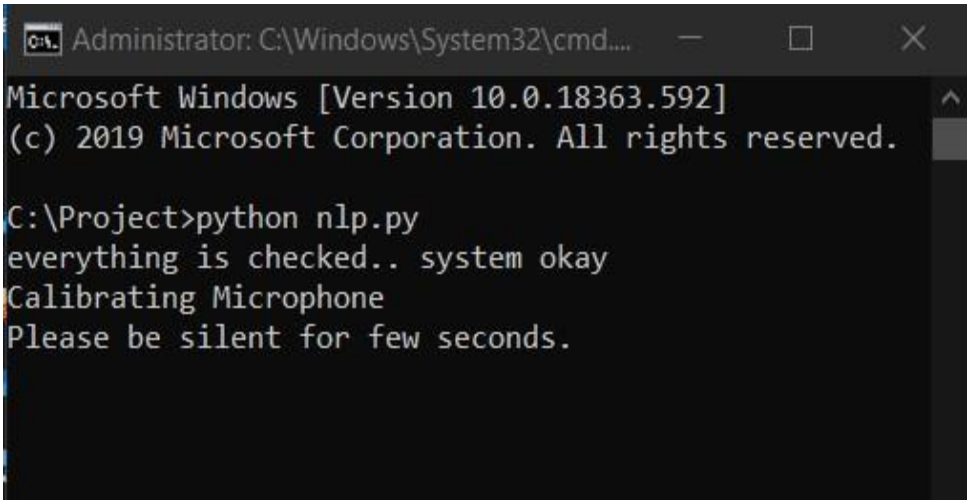

Fig: 3 Execution Process

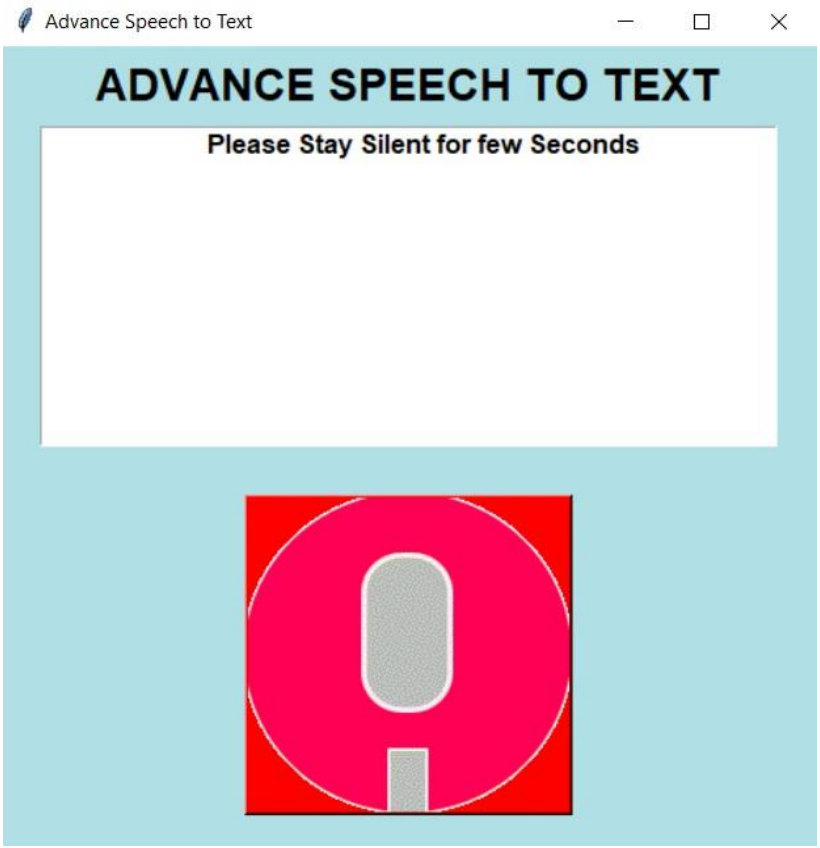

Fig: 4 Instructions on Output Screen

DOI:10.48047/INTJECSE/V15I4.93 ISSN: 1308-5581 Vol 15, Issue 04 2023

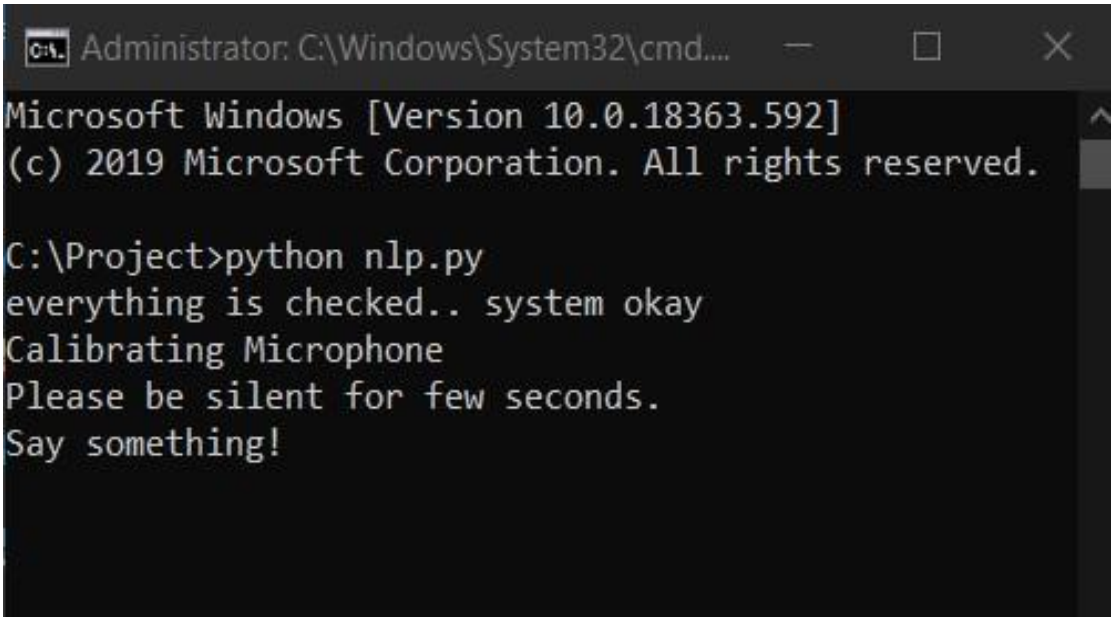

Fig5 Generating Instructions on Command Prompt

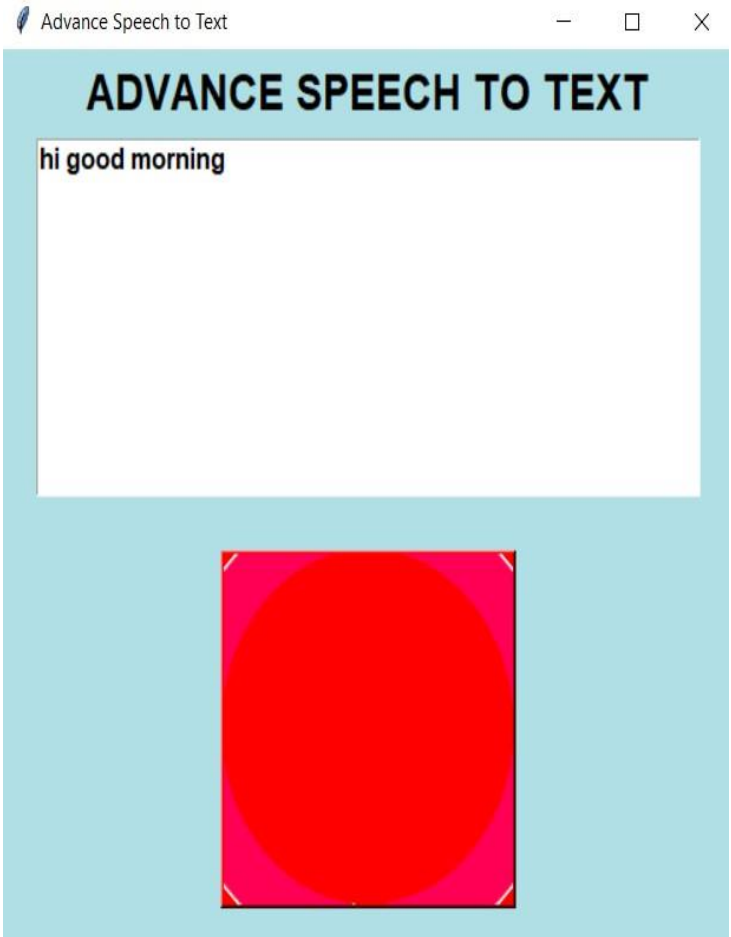

Fig 6. Speech Converted into Text

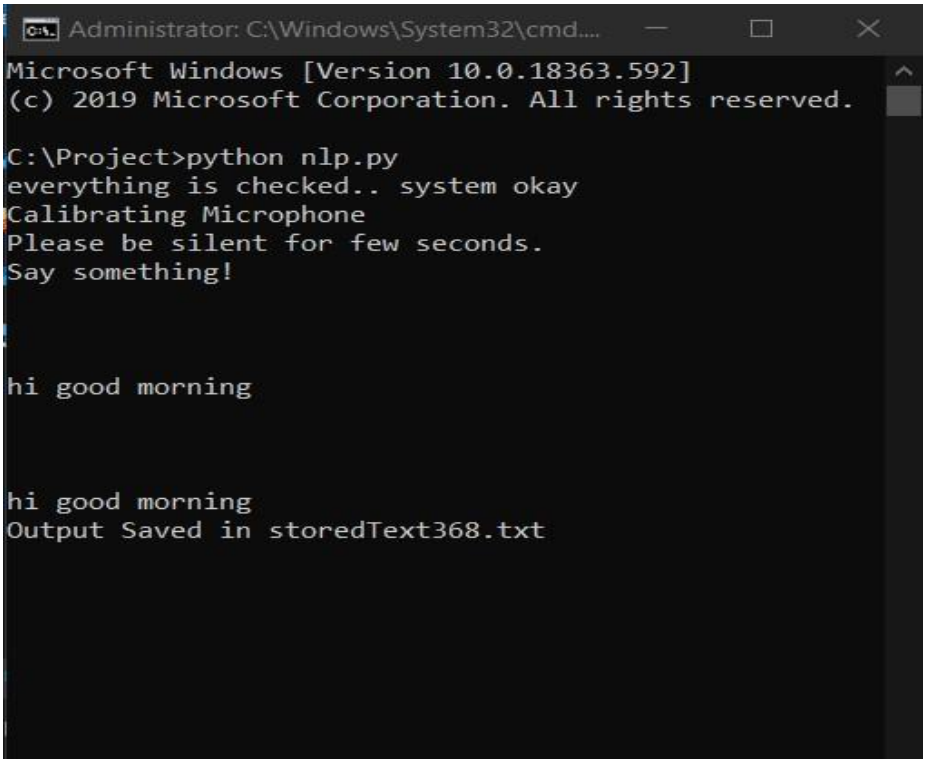

Fig: 7 Output Stored in text file

CONCLUSION

Speech recognition has a big potential in becoming an important factor of interaction between human and machine in the near future. Speech recognition has been developed steadily over the last decades and it has been incorporated into several new applications. The results show reasonably good success in recognizing continuous speech from various speakers. The different modules were analyzed in their respective domains and were successfully verified for different speech inputs. The speaker independent speech recognition systems were successfully trained to recognize speech inputs that were recorded using a microphone. This work can be taken into more detail and more work can be done on the project in order to bring modifications and additional features. The current version of the project support only few areas of the notepad but more areas can be covered and effort will be made in this regard. At some point in the future, speech recognition may become speech understanding. However it requires more computational power to grasp the meaning behind it. Some researchers argue that speech recognition development offers the most direct line from the computers of today to true Artificial Intelligence.

REFERENCES

1. Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition: The shared views of four research groups. IEEE Signal Processing Magazine, 29(6), 82-97.
2. Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. Proceedings of the 23rd International Conference on Machine Learning (ICML), 369-376.
3. Deng, L., & Li, D. (2013). Machine Learning Paradigms for Speech Recognition: An Overview. IEEE Transactions on Audio, Speech, and Language Processing, 21(5), 1060-1089.
4. Hannun, A. Y., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., & Ng, A. Y. (2014). Deep Speech: Scaling up end-to-end Speech Recognition. arXiv preprint arXiv:1412.5567.

5. Graves, A., Mohamed, A. R., & Hinton, G. (2013). Speech Recognition with Deep Recurrent Neural Networks. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6645-6649.
6. Amodei, D., Anubhai, R., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Chen, J., Chrzanowski, M., Coates, A., Diamos, G., et al. (2015). Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. arXiv preprint arXiv:1512.02595.
7. Lippmann, R. P. (1987). An Introduction to Computing with Neural Nets. IEEE ASSP Magazine, 4(2), 4-22.
8. Rabiner, L. R., & Juang, B. H. (1993). Fundamentals of Speech Recognition. Prentice Hall.
9. Young, S., Evermann, G., Gales, M. J. F., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., et al. (2006). The HTK Book (for HTK Version 3.4). Cambridge University Engineering Department.
10. Baker, J. K. (1975). The Dragon System - An Overview. IEEE Transactions on Acoustics, Speech, and Signal Processing, 23(1), 24-29.
11. Huang, X., Acero, A., Hon, H., & Reddy, R. (2001). Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Prentice Hall.
12. Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE, 77(2), 257-286.
13. Jurafsky, D., & Martin, J. H. (2008). Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. Pearson.
14. Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., & Woodland, P. (2002). The HTK Book (for HTK Version 3.2). Cambridge University Engineering Department.
15. Lee, K. F. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE, 77(2), 257-286.
16. Bahl, L. R., Brown, P. F., de Souza, P. V., & Mercer, R. L. (1986). Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 49-52.
17. Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A Neural Probabilistic Language Model. Journal of Machine Learning Research, 3, 1137-1155.
18. Sak, H., Senior, A., & Beaufays, F. (2014). Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling. arXiv preprint arXiv:1402.1128.
19. Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016). Listen, Attend and Spell. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4960-4964.
20. Graves, A., Wayne, G., & Danihelka, I. (2014). Neural Turing Machines. arXiv preprint arXiv:1410.5401.