Speech to Text Conversion with Multiple Indian Languages

Arthat Arora, Gouri Sankar Mishra, Parma Nand, Rani Astya, Pradeep Kumar Mishra, Nikhil Sharma, Anubha Sah, Abhinav Srivastava

Department of Computer Science and Engineering, SET, Sharda University, India

arthatarora@gmail.com, nikhil.k2524@gmail.com, sahanubha12@gmail.com, abhi.s007ea@gmail.com, gourisankar.mishra@sharda.ac.in, parma.nand@sharda.ac.in, rani.astya@sharda.ac.in, pradeepkumar.mishra@sharda.ac.in

Abstract

Communication is one of the vital components to pass on any kind of information and communicate with individuals. Various applications are used for speech recognition, speech to text conversion, language interpretation, and so on. SPEECH TO-TEXT recognition is a product that allows the client to enhance the cognition capacity and directs text by voice. This paper focuses on the speech to text conversion of English to multiple Indian languages by using the methodologies of natural language processing and machine learning. Since a large number of the issues emerging in speech recognition are appropriate for algorithmic examinations, we present them in wording recognizable to algorithm designers.

Keywords: Speech Recognition, Speech to text, Automatic Speech Recognition, Interpretation

1. INTRODUCTION

Natural language processing (NLP) is a branch of artificial intelligence in the computer technological that makes the assisting computers to recognize the manner that human beings write and talk. This is a tough undertaking as it includes numerous unstructured statistics. The fashion in which human beings communicate and write (referred to as 'tone of voice') is unique to individuals, and constantly evolving.

Understanding context is also a difficult task – something that requires semantic analysis for system gaining knowledge of to get a deal with on it. Natural language understanding (NLU) is a sub-branch of NLP and deals with those nuances via gadget analyzing comprehension in preference to without a doubt knowledge literal meaning. The goal of NLP and NLU is to assist computer systems recognize human language properly enough that they can communicate in a natural manner. Some of the common applications of NLP are : Voice Assistants, Different service based chatbots, Language Translation, Voice based application

1.1. Working of NLP

The working of NLP is divided into 5 steps; starting from the lexical analyzer to the pragmatic analyzer. The input speech is sub-divided at each step for the simplification of the speech and converted to small blocks of information that can be processed and understood by the machine to perform the specific task. The 5 stages of analyzing sentence are represented in Fig.1.

The initial step involves lexical analyzer which helps in the simplification of the huge text into small blocks called lexicons, it divides or classify the text into words, sentence, phrases or paragraphs. The processed classification is then moved to syntactic analyzer which checks and arranges the text into correct grammar format and forms a meaning of the sentence that is understood by the user. Syntactic analyzer arranges the words and phrases in such a format that in form a correct readable sentence. The text is moved further to semantic analyzer which perform the task of converting the formed sentence to show case the actual meaning of the sentence i.e checking if the formed sentence is having the same meaning as it is intend to make. Once the

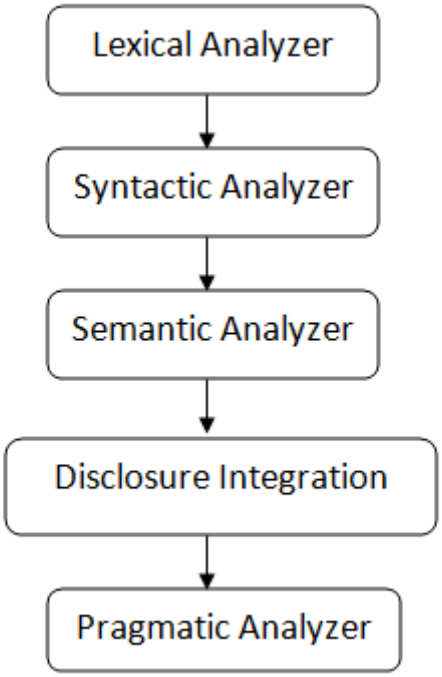

Figure 1: Working Flow of NLP

sentence is verified for its meaning then it moves to the next phase of disclosure integration, which helps in a formation of group of sentences without losing the meaning of each line and converting them into a meaning paragraph. The Disclosure Integration checks for each sentence meaning with the sentence just before it to check if the correct meaningful sentence is maintained. The final stage is the pragmatic analyzer in which the whole output is checked

is it correct and have some relation with the real-world knowledge. After all the successful processing of all the phases the system processes the desired output.

1.2. Automated Speech Recognition (ASR)

Automatic Speech Recognition (ASR) is an innovation or a technology that grants individuals to utilize their voices to talk and have conversation with a PC interface in a way like a typical human discussion. The most progressive adaptation of at present created ASR innovations spins around what is called Natural Language Processing, or NLP. This variation of ASR comes the nearest to permitting genuine discussion among individuals and machine intelligence.

Lately, ASR has become famous in the customer care divisions of huge corporation. It is additionally utilized by a few government offices and different associations. Some of the ASR frameworks perceive the input in a single word like yes/no or numeral which makes it feasible for people to manage computerized options without typing numerals and has no capacity to bear any kind of error. In a manual-section circumstance, a client could hit some unacceptable key subsequent to having entered 20 or 30 numerals at stretches already in the menu, and surrender as opposed to bringing again and beginning once again. ASR for all intents and purposes disposes of this issue. Refined ASR frameworks permit the client to enter direct questions or reactions, for example, a request for driving direction or the phone number of a lodging in a specific town. It likewise decreases the quantity of directions that the client should get and comprehend.

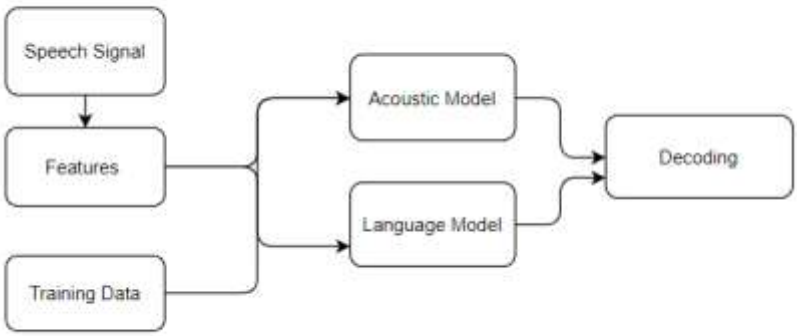

Figure 2: Working of ASR

In the easiest terms, speech recognition happens when a PC gets audio input from an individual talking, processes the input by separating the different parts of speech, & afterward translates that speech to message.

Some ASR frameworks are speaker-dependent and should be prepared to perceive specific words and speech pattern. These are basically the voice-recognition frameworks utilized in your smart devices. You need to say explicit words and expressions into your telephone

before the ASR-powered voice assistant begins working for it to figure out how to recognize your voice.

Other ASR frameworks are speaker-independent. These frameworks don't need any preparation or training. Speak independent frameworks can perceive verbally expressed words regardless of the speaker.

2. SPEECH RECOGNITION-A STATE OF ART

This mixture strategy helps application that requires brief outline of extended speeches which is very valuable for documentation. One of the most important step while working with NLP is to extricate the components of speech having few qualities. It turns into a kind of obstruction to summarization process by supposing a word or a sentence is perceived as negligible. Even punctuation assumes an indispensable part in synopsis as semantics is significant while summing up the content. The methodology proposed by the team was to summarize the text extracted through the input with respect to the rank of the sentences. The frequency of occurrence of words can be used to determine the rank of the sentences. And to find the frequency of words, they used the sentence tokenize and word tokenize techniques are available in python NLTK packages. Using Google API, text is extracted and then the sentences are obtained using sentence tokenize and words are extracted using word tokenize. This input received through the user is converted into signals and then converted into text format.

This aricle on “Speech to Text Conversion Methods” explained varieties of the speech signal and their significance in automatic speech recognition. A database has been made from the different words and syllables. The ideal speech is delivered by the Concatenative speech synthesis methodology. The framework gives the input information as voice, then, preprocessed that information and changed over into text showed on PC. The client types the input string and the framework peruses it from the database or information store where the words, telephones, diaphones, triphones are put away. This system had a speech to text system with the vocabulary of ten words i.e. digits 0 to 9 and statistical modelling (HMM). HMM was used for machine speech recognition. This system builds an HMM model using C programs for each word present in vocabulary. This is done during the training phase. However, during the recognition phase, speech is acquired and stored in FPGA's memory to preprocess and calculate the probability of observation sequence.

The authors proposed a framework named “ScribeBot” to help the visually impaired students to pro their assessments and exams. SCRIBEBOT is a Raspberry Pi 3 coordinated with a headset with Microphone, a Monitor and a printer. The framework they've made uses the advanced deep learning algorithms and neural organizations which have been carried out in the Google speech API. The two principle measures associated with their work is the speech to text transformation and furthermore the way toward changing the content over to speech so the students can think about the question that has been asked in the paper. The framework has been executed utilizing python programming language which imports the Google Speech API package alongside the Google Text To Speech package.

This study had talked regarding the strategies of dynamic time traveling and mel scale repeat cepstral constant within the confined talk affirmation. Unique elements of the

communicated word have been freed from the information talk. Associate degree illustration of five speakers has been assembled and each one of them had spoken ten digits. A information set is formed on this premise. Then, at that time highlight has been separated utilizing MFCC. DTW is used for adequately overseeing entirely unexpected talking speed. It is used for closeness assessment between 2 plans that shifts in speed & time.

Table 1. A comparative summary of existing related surveys

Author(s)	Year	Technique
"Vinnarasu A., Deepa V. Jose"	2019	Use of NLP and text summarization by neglecting repeated words.
"Dhanush Kumar S, Lavanya S, Madhumita G and Mercy Rajaselvi V"	2018	ScribeBot: Made with Raspberry Pie and Google API
"Mittal et al."	2018	ASR framework for Punjabi language under various acoustic conditions(created explicitly for mobile phones)
"Sagar Patil, Mayuri Phonde, Siddharth Prajapati, Saranga Rane and Anita Lahane"	2016	ASR framework for Punjabi language under various acoustic conditions(created explicitly for mobile phones)
"Miss. Prachi Khilari and Prof. Bhope V. P."	2015	Database of different words and syllables for conversion of speech to show text on monitor by concentrating on string rather than information.
"Geeta Nijhawan, Poonam Pandit and Shivanker Dev Dhingra"	2013	Strategies of dynamic time traveling and mel scale repeat cepstral constant(Sub band centroid)
"Puneet Kaur, Bhupender Singh and Neha Kapur"	2012	Recognizing speech by Hidden Markov Model
"Jingdong Chen and et al"	2004	mel-recurrence cepstral coefficients (MFCCs) in clean discourse, while passing preferred execution over MFCC in boisterous conditions.

The author examined that notwithstanding their boundless fame as frontend boundaries for discourse acknowledgment, the cepstral coefficients which has been received from either

direct assumption examination or a channel bank are found to be tricky to added substance clamor. Here in this letter, we examine the utilization of ghastly sub band centroids for good discourse acknowledgment. We show that centroids, assuming appropriately chosen, can achieve acknowledgment execution tantamount to that of the mel-recurrence cepstral coefficients (MFCCs) in clean discourse, while passing preferred execution over MFCC in boisterous conditions. A technique is proposed to assemble the unique centroid highlight vector that essentially typifies the temporary apparition information.

The author proposed and executed an ASR framework for Punjabi language under various acoustic conditions. The significant limit of setting subordinate unfastened model is the necessity of higher memory space. This framework is explicitly created for cell phones.

The author had discussed a way to use Hidden Markov Model in the method of popularity of speech. The crucial 3 steps that are important to broaden an 'Automatic Speech Recognition' machine are pre-processing, characteristic Extraction and popularity and subsequently hidden markov version is used to get the required end result. Since there are already huge amount of improvements inside the discipline of virtual sign processing, research persons are trying their best to broaden a ideal ASR device. However, the overall performance of PCs within identical timings aren't that high in terms of matching speed and accuracy.

The author proposed a framework named "Multilingual Speech and Text Recognition and Translation using Image" to robotize the application to defeat from the language boundary in between the nations & furthermore states inside the country. They carried out framework for client who staging issues of language obstruction and furthermore its user interface is additionally easy to use so that the client can undoubtedly collaborate with this framework. So due to this framework, users don't need to utilize word reference for knowing the importance of word, hence it naturally decreases the client task for knowing the language for communication.

3. PROPOSED METHODOLOGY

The system works with the understanding or the audio signal of stream or audio input through the microphone jack and interpretation of the signals into the data is done by pre-requisite audio files which can understand the audio inputs and act upon them. There is a database same as on which Google API works which can understand the spoken words and get the correct grammar and meaning and form the correct meaningful sentence.

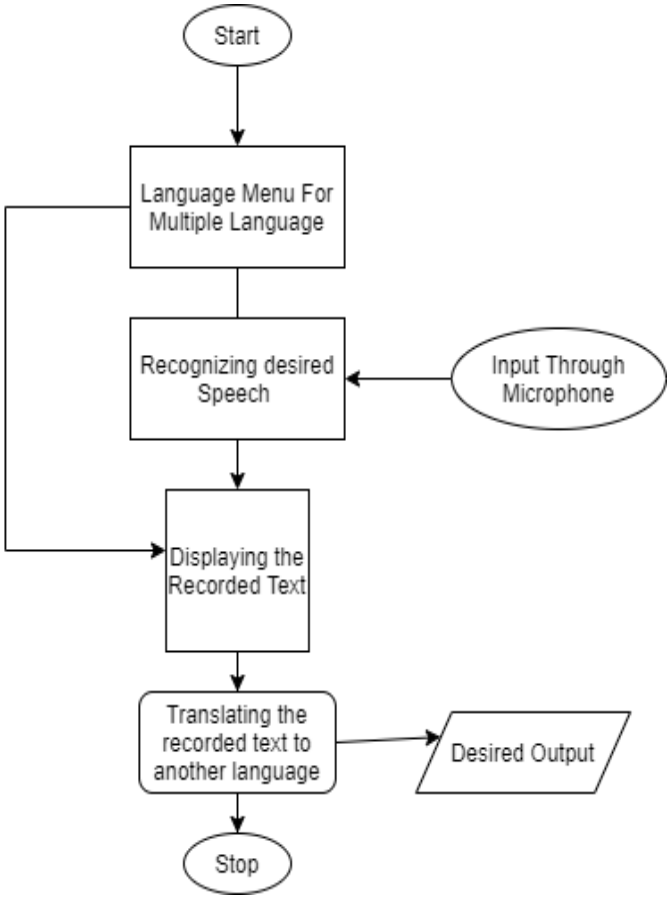

Figure 3: Process Flowchart

The successful sentence formed is then processed for further evaluation and conversion, the system work both for multi-language support to remove to banner of knowledge of all the language for the communication around the globe. The processed data in the user language is then converted to desired client language for understanding of what another user is saying. The working is as follow:

- There is a menu of multiple language support for conversion which user can use to translate its text and record in the selected desired language.
- After the selection the user inputs the data as an audio signal through the microphone.
- The recorded input text is showing to the user for verification if it is actual correct to what is being said.
- The menu is again shown to the user to convert the recorded text to desired language that he wants to translate.
- After successful processing, the desired output is given of the screen after conversion for further use.

The main advantage of this system is the simplicity of communication and speedy document turnaround. It helps us to perform multiple tasks while dictating and can create records in under a fraction of the time it takes to type. It provides us the adaptability to work in or out of the workplace. Also, the time is saved with expanded productivity and has less desk work.

The main disadvantage of this system is the absence of accuracy and misinterpretation. In the event that you talk excessively quick or vaguely, you'll increment spelling and punctuation blunders. It will most likely be unable to separate between your speech, others talking and other surrounding noise, prompting record misunderstandings and mistakes.

he functions of system is to provide the language option for the user in the language menu, receive audio stream through microphone in the form of speech, classify audio signal addressee at the run time, provide the desired output in the preferred language selected by the user, map applicable signal to word series and then to action request, and user feedback.

4. CONCLUSION

Speech Recognition is a vast area to explore. Speech recognition framework permits PCs to take spoken sound, interpret it and create text from it. The speech to text conversion might appear to be compelling and effective to its clients in the event that it produces regular speech and by making a few alterations to it. Our proposed system allows the user to recognize their speech and convert them into text in English as well as some Indian Languages provided in the option. This helps the user who find themselves uncomfortable in English language and provides leverage to have their text converted into their preferable language.

REFERENCES

[1] Vinnarasu A., Deepa V. Jose, “Speech to text conversion and summarization for effective understanding and documentation”, IOSR Journal of Computer Engineering (IOSR-JCE), Vol. 9, No. 5, October 2019.

[2] Miss.Prachi Khilari, Prof. Bhope V. P., “A REVIEW ON SPEECH TO TEXT CONVERSION METHODS”, IJARIIT, Vol. 4, July 2015.

[3] Dhanush Kumar S, Lavanya S, Madhumita G, Mercy Rajaselvi, “Journal of Speech to Text Conversion”, IJARCET, Vol. 4.

[4] Shivanker Dev Dhingra, Geeta Nijhawan, Poonam Pandit, “Isolated Speech Recognition using MFCC and DTW” International Journal of Advance Research in Electrical, Electronics and Instrumentation Engineering, Vol.2, Issue 8, August 2013.

[5] Jingdong Chen, Member, Yiteng (Arden) Huang, Qi Li, Kuldip K. Paliwal “Recognition of Noisy Speech Using Dynamic Spectral Subband Centroids” in IEEE SIGNAL PROCESSING LETTERS, VOL. 11, NO. 2, FEBRUARY 2004

[6] Ayushi Trivedi,Navya Pant, Pinal Shah,Simran Sonik and Supriya Agrawal, “Speech to text and text to speech recognition systems-Areview”, IOSR Journal of Computer Engineering (IOSR-JCE), Volume 20, Issue 2, Ver. I (Mar.- Apr. 2018)

[7] Bhupinder Singh, Neha Kapur, Puneet Kaur “Sppech Recognition with Hidden Markov Model:A Review” International Journal of Advanced Research in Computer and Software Engineering, Vol. 2, Issue 3, March 2012.

[8] Sagar Patil, Mayuri Phonde, Siddharth Prajapati, Saranga Rane and Anita Lahane, "Multilingual Speech and Text Recognition and Translation using Image", International Journal of Engineering Research & Technology (IJERT), Vol. 5 Issue 04, April-2016

[9] Shaikh Naziya S.1*, R.R. Deshmukh2, "Speech Recognition System – A Review", IOSR Journal of Computer Engineering (IOSR-JCE), Volume 18, Issue 4, Ver. II (Jul.-Aug. 2016)

[10] Anchal Katyal, Amanpreet Kaur, Jasmeen Gill, "Automatic Speech Recognition: A Review", International Journal of Engineering and Advanced Technology (IJEAT), Volume-3, Issue-3, February 2014

[11] Manjutha M, Gracy J, Dr P Subhashini, Dr M Krishnaveni, "Automated Speech Recognition System – A Literature Review", International Journal of Engineering Trends and Applications (IJETA) – Volume 4 Issue 2, Mar-Apr 2017

[12] T.Y.G, "Development and Comparison of ASR Models using Kaldi for Noisy and Enhanced Kannada Speech Data," pp. 1832–1838, 2017.

[13] P. Mittal and N. Singh, "Development and analysis of Punjabi ASR system for mobile phones under different acoustic models," Int. J. Speech Technol., vol. 0, no. 0, p. , 2019.

[14] S. Huang and S. Renals, "Hierarchical Bayesian Language Models for Conversational Speech Recognition," vol. 18, no. 8, pp. 1941–1954, 2010.

[15] Sarika Hegde ,K K Achary,Surendra Shetty- Analysis of Isolated Word Recognition for Kannada Language using Pattern Recognition Approach-International Journal of Information Processing , 7(1), 73 - 80. -2013.

[16] Sai Prasad P. S. V. S, . Girija P. N Speech Recognition of Isolated Telugu Vowels Using Neural Networks. Proceedings of the 1st Indian International Conference on Artificial Intelligence, IICAI 2003, Hyderabad, India, 28-34 December 18-20, 2003.

[17] Hegde, R.M., Murthy, H.A., Gadde, V.R.R., 2004. Continuous speech recognition using joint features derived from the modified group delay function and MFCC. In: Proceedings of ICSLP, Jeju, Korea, pp. 905–908.

[18] Hegde, R.M., Murthy, H.A., Rao, G.V.R., 2005. Speech processing using joint features derived from the modified group delay function. In:Proceedings of ICASSP, vol. I. Philadelphia, PA, pp. 541–544.

[19] Sunitha .K.V.N & Kalyani.N (2009). Syllable analysis to build a dictation system in Telugu language. International Journal of Computer Science and Information Security. Vol 30. No 30.

[20] Sunitha K V N, Kalyani N. Isolated Word Recognition using Morph Knowledge for Telugu Language. International Journal of Computer Applications 38(12):47-54, February 2012. Published by Foundation of Computer Science, New York, USA.

[21] Vijai Bhaskar P , Rama Mohan Rao S , Gopi A - "HTK Based Telugu Speech Recognition" International Journal of Advanced Research in Computer Science and Software Engineering - Volume 2, Issue 12, December 2012.

[22] Usha Rani N, Girija P N "Error analysis to improve the speech recognition accuracy on Telugu language" Sadhana Vol. 37, Part 6, December 2012, pp. 747–761. Indian Academy of Sciences.

[23] Himanshu N. Patel, P.V. Virparia - "A Small Vocabulary Speech Recognition for Gujarati"- International Journal of Advanced Research in Computer Science Volume 2, Issue 1, Jan-Feb 2011. ISSN 0976-5697.

[24] Patel Pravin, Harikrishna Jethva -" Neural Network Based Gujarati Language Speech Recognition " International Journal of Computer Science and Management Research Vol 2 Issue 5 May 2013.

[25] M.K.Deka , C.K.Nath , S.K.Sarma , P.H. Talukdar - "An Approach to Noise Robust Speech Recognition using LPC-Cepstral Coefficient and MLP based Artificial Neural Network with respect to Assamese and Bodo Language" International Symposium on Devices MEMS, Intelligent Systems & Communication (ISDMISC) 2011.

[26] Utpal Bhattacharjee-"Recognition of the Tonal Words of BODO Language"- International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-1, Issue-6, January 2013.

[27] Syama R, Suma Mary Idikkula (2008) "HMM Based Speech Recognition System for Malayalam", ICAI'08 – The 2008 International Conference on Artificial Intelligence, Monte Carlo Resort, Las Vegas, Nevada, USA (July 14-17, 2008) .

[28] Krishnan, V.R.V. Jayakumar A, Anto P B (2008) , "Speech Recognition of isolated Malayalam Words Using Wavlet features and Artificial Neural Network". DELTA2008. 4th IEEE International Symposium on Electronic Design, Test and Applications, 2008.Volume,Issue, 23-25 Jan. 2008 Page(s):240 – 243.

[29] A.R. Sukumar, A.F. Shah, and P.B. Anto, "Isolated question words recognition from speech queries by using Artificial Neural Networks", in proc. of IEEE 2nd International conference on Computing, Communication and Networking Technologies(ICCCNT), Karur, India, 2010, pp. 1-4.

[30] Anuj Mohamed, K. N. Ramachandran Nair-Continuous Malayalam speech recognition using Hidden Markov Models. Proceedings of the 1st Amrita ACM-W Celebration on Women in Computing in India, September 16-17, 2010, Tamilnadu, India; 01/2010.

[31] Sonia Sunny, David Peter S , and K Poulose Jacob -A Comparative Study of Wavelet Based Feature Extraction Techniques in Recognizing Isolated Spoken Words," International Journal of Signal Processing Systems, Vol.1, No.1, pp.49-53,June 2013.

[32] Kavita Sharma, Prateek Hakar "Speech Denoising Using Different Types of Filters" International journal of Engineering Research and Applications Vol. 2, Issue 1, Jan-Feb 2012

[33] Om Prakash Prabhakar, Navneet Kumar Sahu,"A Survey on Voice Command Recognition Technique" International Journal of Advanced Research in Computer and Software Engineering, Vol 3,Issue 5,May 2013.

[34] Shivanker Dev Dhingra, Geeta Nijhawan, Poonam Pandit, "Isolated Speech Recognition using MFCC and DTW" International Journal of Advance Research in Electrical, Electronics and Instrumentation Engineering, Vol.2, Issue 8, August 2013.

[35] Hakan Erdogan, Ruhi Sarikaya, Yuqing Gao, "Using semantic analysis to improve speech recognition performance" Computer Speech and Language, ELSEVIER 2005.

[36] Ibrahim Patel, Dr. Y. Srinivas Rao, "Speech Recognition using HMM with MFCC-an analysis using Frequency Spectral Decomposition Technique" Signal and Image Processing:An International Journal(SIPIJ),s Vol.1, Number.2, December 2010.