International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 12 | Dec 2020 www.irjet.net p-ISSN: 2395-0072

Interactive Speech Recognition Agent System using AI

Akshata Tayade[1], Mahima Thakur[2], Laxmi Vathari[3], Prof. Satish Kuchiwale[4]

[1,2,3]Student, Dept. of Computer Engineering, Smt. Indira Gandhi College of Engineering, Navi Mumbai, Maharashtra, India
[4]Professor, Dept. of Computer Engineering, Smt. Indira Gandhi College of Engineering, Navi Mumbai, Maharashtra, India

---

Abstract - *To assist the disabled we may make use of speech to text translation technique, speech recognition technology that turns spoken words into written words. It can also identify and understand human speech to carry out a person's command on an android phone. This paper will help everyone to communicate with one another, it will enable user to interact with everyone. Basically it will convert audio signal which we can hear on our phone into text and then the text can be used for documentation or any other work and for call the conversation can be saved which is in the form of text.*

***Key Words*: ASR - Automatic speech Recognition, Dictation- In which the user enters the data by reading directly to the computer, STT-Speech to text**

1. INTRODUCTION

Voice is the basic, common, and efficient form of communication method for people to interact with each other. Today speech technologies are commonly available for a limited but interesting, various range of task. This technology enables machines to respond accurately and reliably to human voices and provide useful and valuable services. Communicating with a computer is faster using voice rather than using keyboard, so people will prefer such a system. Communication between the human being is dominated by spoken language, therefore it is natural for people to expect voice interfaces with computer. This can be accomplished by developing a voice recognition system: speech-to-text which allows a computer to translate voice request and dictation into text.

The supposed system is an application that converts speech into text with the help of emerging technology Neural network, Artificial intelligence, Deep learning, Machine learning. As in the world of digitalization, it needs to build such a system to solution to some circumstances in education, communication, and in daily life. It can identify and understand human speech and convert it into the form of text. Our app will have different models to work for different features and will be having an Automatic Speech Recognition System (ASR). The Speech to Text App will be used in many aspects for the betterment of the people in bunch of features such as Generation of subtitle on the screen for ongoing audio on your device, Language Translation[2], Call to text conversion[4] and video captions from an online source as URL link i.e. Subtitle for video on internet. So now he/she can learn new things without any kind of obstacle in the studies. There will be less communication and education barrier and ease of day to day life for deaf and old aged people.

1.1. PROBLEM STATEMENT

To design an automatic speech recognition system that gives the best recognition results for both male and female speakers. Such that, it can identify and understand human speech and convert it into the form of text. The subtitle for any video on the internet will be provided by supposed system if you provide file in app to lessen the linguistic barrier in education. If the other end caller is speaking fast and the user is not able to understand then the user can convert the audio call into text. The user get text captions for particular audio playing on screen as it may be a lecture for online classes or the any video which is in any other language. It will also give solution to the person who is traveling alone in unknown country to communicate, all these features will be provided by supposed system.

1.2. OBJECTIVES

1) Generation of subtitle on the screen for ongoing audio on your device.
2) It destroys language barrier present in communication.
3) Subtitle for video on internet.
4) Call conversion to text will be available.
5) Language Translation in English.

1.3. SCOPE

1) Speech to text conversion becomes a link between deaf and normal person. It will helps the person with a hearing disability by providing subtitle to any video available on the internet.
2) It will let the user having hearing disability to answer calls. The app may help them to know what the other end caller is saying by reading the subtitles which will get converted by the end caller's speech.

3) Also, the one, who wants to understand any kind of video in one of particular language by translating the language through this app.

4) And the person who don't have earphones or cannot hear it properly, can use this application to get the subtitles.

2. ANALYSIS

2.1. SURVEY ANALYSIS

Esfandier Zavarehei and et al in the year 2005, studied that a time-frequency estimator for enhancement of noisy speech signal in DFT domain is introduced. It is based on low order auto regressive process which is used for modelling. The time-varying trajectory of DFT component in speech which has been formed in Kalman filter state equation. For restarting Kalman filter, a method has been formed to make alteration on the onsets of speech. The performance of this method was compared with parametric spectral substraction and MMSE estimator for the increment of noisy speech. The resultant of the proposed method is that residual noise is reduced and quality of speech in improved using Kalman filters [2].

Ibrahim Patel and et al in the year 2010, had discussed that frequency spectral information with mel frequency is used to present as an approach in the recognition of speech for improvement of speech, based on recognition approach which is represented in HMM. A combination of frequency spectral information in the conventional Mel spectrum which is based on the approach of speech recognition. The approach of Mel frequency utilize the frequency observation in speech within a given resolution resulting in the overlapping of resolution feature which results in the limit of recognition. In speech recognition system which is based on HMM, resolution decomposition is used with a mapping approach in a separating frequency. The result of the study is that there is an improvement in quality metrics of speech recognition with respect to the computational time and learning accuracy in speech recognition system[6].

Kavita Sharma and Prateek Hakar in the year 2012 has represented recognition of speech in a broader solutions. It refers to the technology that will recognize the speech without being targeted at single speaker. Variability in speech pattern, in speech recognition is the main problem. Speaker characteristics which include accent, noise and co-articulation are the most challenging sources in the variation of speech. In speech recognition system, the function of basilar membrane is copied in the front-end of the filter bank. To obtain better recognition results it is believed that the band subdivision is closer to the human perception. In speech recognition system the filter which is constructed for speech recognition is estimated of noise and clean speech[10].

Geeta Nijhawan, Poonam Pandit and Shivanker Dev Dhingra in the year 2013 had discussed the techniques of dynamic time warping and mel scale frequency cepstral coefficient in the isolated speech recognition. Different features of the spoken word had been extracted from the input speech. A sample of 5 speakers has been collected and each had spoken 10 digits. A database is made on this basis. Then feature has been extracted using MFCC.DTW is used for effectively dealing with various speaking speed. It is used for similarity measurement between two sequence which varies in speed and time[5].

2.2. METHODOLOGY

As an emerging technology, not all developers are familiar with speech recognition technology. While the basic functions of both speech synthesis and speech recognition takes only few minutes to understand (there are subtle and powerful capabilities provided by computerized speech that developers will want to understand and utilize. An understanding of the capabilities and limitations of speech technology is also important for developers in making decisions about whether a particular application will benefit from the use of speech input and output.

Speech-Text recognition process

Acoustic model: An acoustic model is a file that contains statistical representations of each of the distinct sounds that makes up a word. Each of these statistical representations is assigned a label called a phoneme. The English language has about 40 distinct sounds that are useful for speech recognition, and thus we have 40 different phonemes.

Language model: Even though the audio clip may not be grammatically perfect or have skipped words, we still assume our audio clip is grammatically and semantically sound. Therefore, if we include a language model in decoding, we can improve the accuracy of ASR.

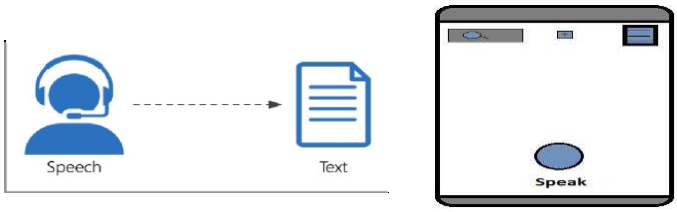

Fig -1: Speech to Text

2.3. DATA FLOW DIAGRAM

A data flow diagram (DFD) is a graphical representation of the "flow" of data through an information system, modelling its process aspects. A DFD is often used as a preliminary step to create an overview of the system, which can later be elaborated.
Following are the given data flow diagrams levels of speech recognition.

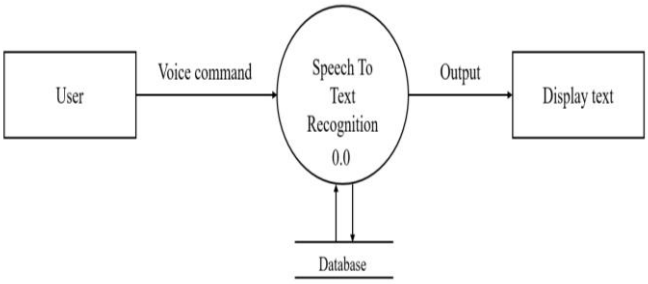

Fig -2: DFD Level 0

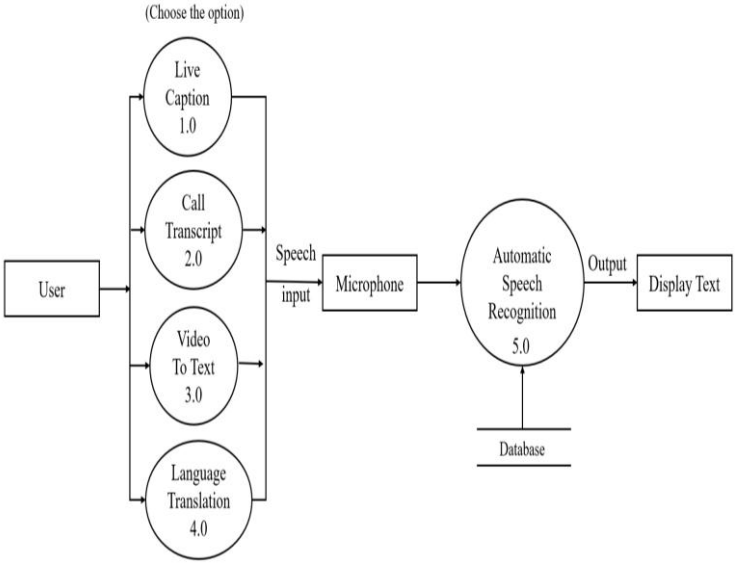

Fig -3: DFD Level 1

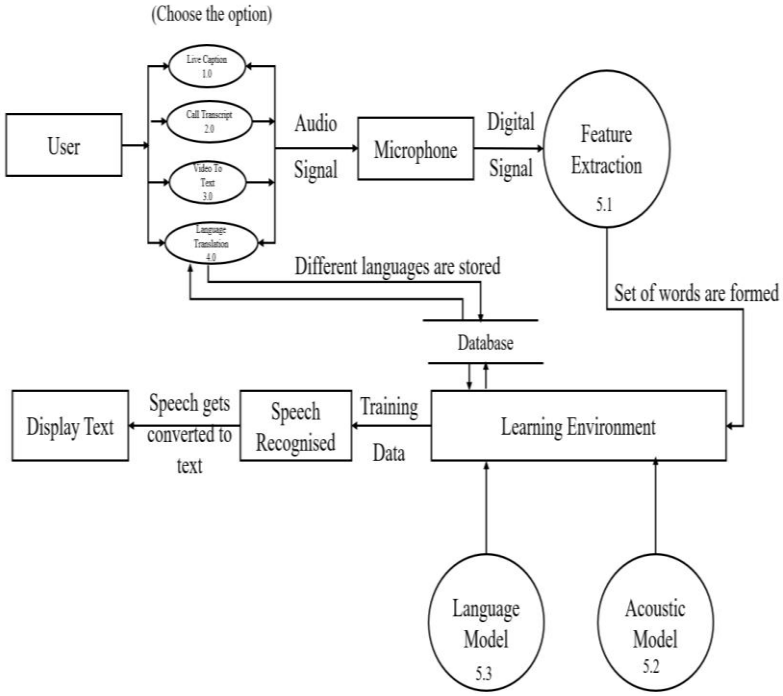

Fig -4: DFD Level 2

3. DESIGN

3.1. FLOWCHART OF COMPLETE SYSTEM

Following figure shows complete system process of speech recognition:

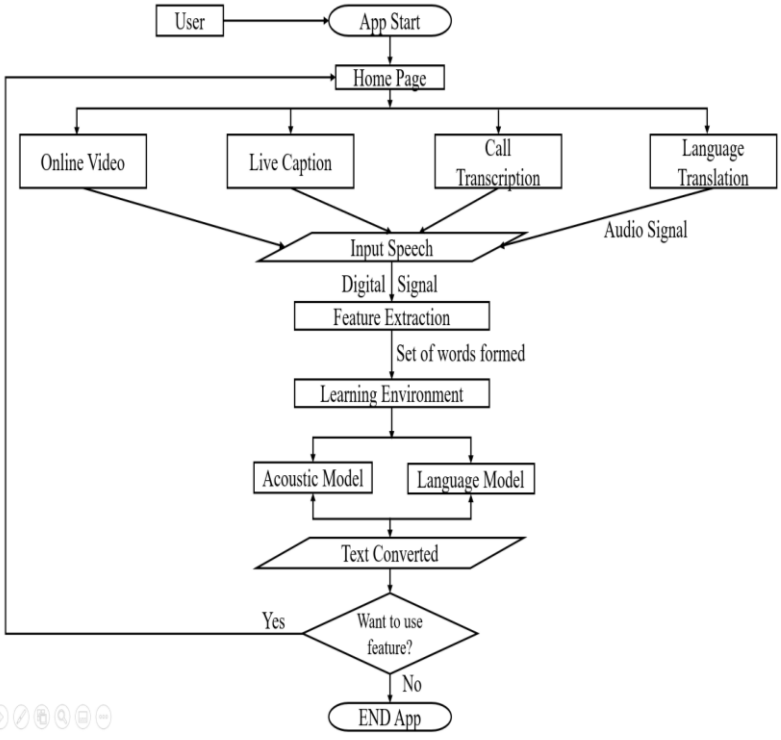

Fig -5: Flowchart of System

3.2. MODULES AND FUNCTIONS

Voice Recognition: For intelligent voice assistant application is done using Server. This process involves the conversion of acoustic speech into a set of words and is performed by software component. Accuracy of speech recognition systems differ in vocabulary size and confusability, modality of speech (isolated, discontinuous, or continuous speech, read or spontaneous speech), task and language constraints. The system consists of five modules: feature extraction, phone model training, dictionary preparation, grammar estimation, and sentence decoding.

Voice Input Manager: It manages the command given by user. It sends the Input given by user to the database manager, Database Manager: It compares Input given by user that is in the form of voice with the database which contains vocabulary of words. It sends response to the action performer

Action Performer: It takes response from the database manager as Input and decides which action should be performed. Action can be in the form of text message or call.

Calling service: The application should allow the users to make a call to the person in the contacts or by saying mobile number of the person to whom user wants to call. By giving a correct command with the calling request to a stored person, the Android phone should successfully direct to the number of the person requested.

Language Translator: It allows user to translate the input speech into Hindi, English. So that the communication between the two people belonging to two different religions.

Pattern Match: Even computer don't understand whatever we are saying but it can recognize using phoneme and can generate basic building blocks so we can use them to form a full sentences using these characters or the words generated by our system and our vocabulary database will correct the grammatical mistake.

3.3. ALGORITHMS

SPEECH TO TEXT

1. Signal level- Feature extraction

- CNN using deep learning
  Since focus on speech recognition based on deep convolutional neural networks (DCNNs). Specifically, we propose a method based on DCNN, which extracts informative features from the speech signal, and those features are then used by Acoustic model for speech recognition. The CNNs are a special variant of conventional feed-forward deep neural networks (DNNs), and have been used in many speech applications.

- MFCC
  The MFCC feature extraction technique basically 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.

2. Acoustic Model

- Connectionist Temporal Classification[(a)- CTC acoustic model]
  It is mainly used in Speech recognition system to detect phoneme present in the audio signal given as input to the convert it to word or we can say labels. Basically it will take sequences of observations as input and work it out on those sequence and it will remove the blank spaces present due to various factors such as utterance of speaker and type of speech whether it is connected speech discontinuous etc. and gives us sequence labels as output.

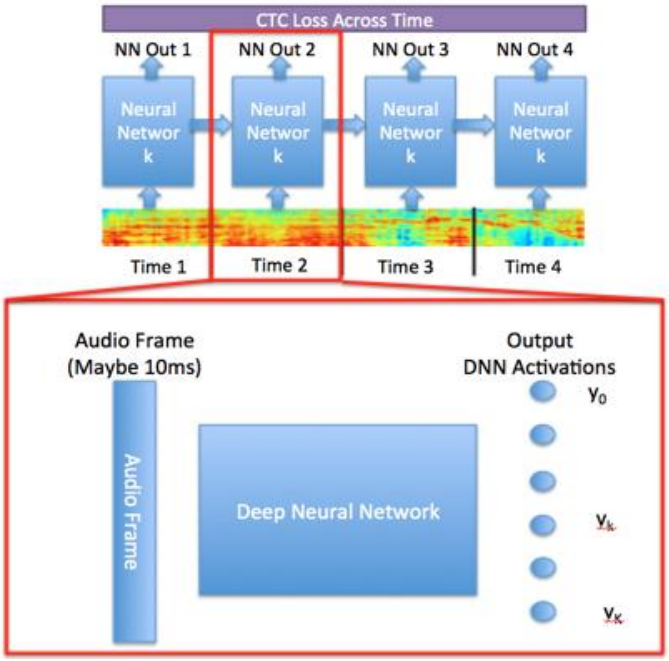

Fig -6: CTC loss (purple) across several NN (blue) outputs Each frame prediction is y.

Use of CTC:- Labelling unsegmented data sequence with recurrent neural network.

- Deep Neural Network

- Convolutional Acoustic model
  Acoustic model in used CNN with one convolution layer, one pooling layer, and 2 fully connected layers. The details of each layer are described in Table 1. The complete topology is shown in Table 1.

Table -1: Details of each CNN layer

Layer	Description
Convolution	• Number of filters: 150 • Filter size: 8 • Stride: 2
Pooling	• Pool size: 6 • Stride: 2
Hidden	• Number of neurons: 1024 • Activation function: ReLU
Output	• Number of neurons: number of states in HMM • Activation function: softmax

It takes output of the feature extraction and then they are passed through convolutional neural network and then given to gated linear units and the dropout units so that we can get character sequence and the these sequences are get into auto-segmented criterion (SG=letters| Waves)which is similar to CTC that removes the duplicates and blanks present in the character sequence And then outputs the words.

3. *Language Model*

- Deep learning
- Recurrent neural network [(b)-CTC Language Model].
  CTC architecture incorporates letter-based language model. CTC architecture can also incorporate a word-based language model by using letter-to-word finite state transducer during decoding. Here we will use TC decoder with LM implementations.
- Convolutional Neural network

LANGUAGE TRANSLATION

- Rule Based Machine Translation (RBMT)
  Translation is generated on the basis of morphological, syntactic, and semantic analysis of both the source and the target languages. Such a system consist of collection of rules:

Grammar rules- basically consist of analysis of languages in terms of grammar structures (syntax, semantic, morphology, part of speech tagging and orthographic features); bilingual or multilingual lexicon dictionary for looking up words during translation while the software program allows effective and efficient interaction of components; and software programs to understand a process those rules. There are three types of rule-based model:

  - **Direct:** It is dictionary based.
  - **Transfer:** It uses lexicons and structural analysis into every SL input text after which it's converted to intermediate representation.
  - **Interlingual:** source language is transformed into an intermediate language which is independent of any of the languages involved in the translation.

- Example based machine translation (EBMT):
  It is based on the idea of analogy. In this approach, the corpus that is used is one that contains texts that have already been translated. Given a sentence that is to be translated, sentences from this corpus are selected that contain similar sub-sentential components. The similar sentences are then used to translate the sub-sentential components of the original sentence into the target language, and these phrases are put together to form a complete translation. The Analogy translation uses three stages; matching, adaption and recombination

  - **Matching**-The SL input text is fragmented, followed by search for examples from database which closely matches the input SL fragment string and the relevant fragments are picked. The TL fragments corresponding to the relevant fragments are extracted.
  - **Adaption**-If the match is exact, the fragments are recombined to form TL output, else find the TL portion of the relevant match correspond to specific portion in SL and align them.
  - **Recombination-** Combination of relevant TL fragments in order to form legal grammatical target text.

4. CONCLUSIONS

The main aim of our project is to develop a system that will allow input voice to see into text format. Communication among the human being is dominated by spoken language, therefore it is natural for people to expect voice interfaces with computer. This can be accomplished by developing voice recognition system: speech-to-text which allows computer to translate voice request and dictation into text. This project made a clear and simple overview of working of speech to text system (STT) in step by step process. The system gives the input data from mice in the form of voice, then preprocessed that data & converted into text format displayed on mobile.

REFERENCES

[1] Jingdong Chen, Member, Yiteng (Arden) Huang, Qi Li, Kuldip K. Paliwal, "Recognition of Noisy Speech using Dynamic Spectral Subband Centroids" IEEE SSIGNAL PROCESSING LETTERS, Vol. 11, Number 2, February 2004.

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

[3] Chadawan Ittichaichareon, Patiyuth Pramkeaw, "Improving MFCC-based Speech Classification with FIR Filter" International Conference on Computer Graphics, Simulation and Modelling (ICGSM'2012) July 28-29, 2012 Pattaya(Thailand).

[4] Dua, M., Aggarwal, R. K., Kadyan, V. and Dua, S. 2012. Punjabi Automatic Speech Recognition Using HTK. IJCSI Int. J. Comp. Sci. 1: 359-364.

[5] 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.

[6] 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), Vol.1, Number.2, December 2010.

[7] 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), Vol.1, Number.2, December 2010.

[8] M A Anusuya, "Speech recognition by Machine", International Journal of Computer Science and Information security, Vol. 6, number 3,2009.

[9] Sikha Gupta, Jafreezal Jaafar, Wan Fatimah wan Ahmad, Arpit Bansal, "Feature Extraction Using MFCC" Signal & Image Processing:An International Journal, Vol 4, No. 4, August 2013.

[10] 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

[11] Mishra, A. N., Astik B. and Mahesh C. 2010. Isolated Hindi Digits Recognition: A Comparative Study. Int. J. Electronics Comm. Eng. 3: 229-238.

[12] Limkar, M., Rama R. and Vidya S. 2012. Isolated Digit Recognition Using MFCC and DTW. Int. J. Adv. Electrical and Electronics Eng. 2: 11-20.

[13] Ye-Yi, W., Acero, A. and Ciprian, C. 2003. Is word Error rate a good indicator for spoken language understanding accuracy. IEEE Workshop on Automatic Speech Recognition and Understanding, St. Thomas, U.S. Virgin Islands, 23–30.

[14] Sakoe, H. and Chiba, S. 1978. Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech and Signal Processing. 26: 43–49.

[15] Baum, L. E. and Petrie, T. 1966. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. The Annals of Mathematical Statistics. 37: 1554–1563

BIOGRAPHIES

"Akshata D. Tayade Student Of Smt. Indira Gandhi College of Engineering, in Computer Science branch, 2017-2021 "

"Mahima H. Thakur Student Of Smt. Indira Gandhi College of Engineering, in Computer Science branch, 2017-2021 "

"Laxmi S. vathari Student Of Smt. Indira Gandhi College of Engineering, in Computer Science branch, 2017-2021 "

"Prof. Satish Lalasaheb Kuchiwale is working as an Assistant Professor in Computer Engineering department in Smt. Indira Gandhi College of Engineering, Ghansoli, Navi Mumbai, affiliated to Mumbai University and having about 13 yrs. of experience. He has completed his M.E. in Computer Engineering."