IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 1, January 2021

DOI 10.17148/IJARCCE.2021.10141

# Dictation Module Using Automatic Speech Recognition in Machine Learning

**Vaishnavi Kocheta[1], Shubhangi Shinde[2], Kiran Nadkar[3], Snehal Jambhulkar[4]**

Student, Department of Computer Engineering, Trinity College of Engineering and Research, Pune, India[1,2,3,4]

**Abstract:** Using Artificial Intelligence and Machine Learning our project proposes a system named Dictation Module. It is a device that can help society to ace their work and help people in multitasking. As it is rightly said that "If a person cannot perform multiple tasks at the same time, he cannot achieve everything he wants". Multitasking is the basic asset of every individual in the society. Unlike others, the socially active businessmen or lawyers are forced to be dependent on this aspect. So we have proposed a device that helps to do multiple tasks anywhere you want. Imagine you are driving and you realize you forgot to mail someone something very important, how will you do it in the traditional way is to reach out to the destination and then start typing your mail and send it, which will cost you lots of time as well as some mishap can occur due to this process. But using our module you can dictate the mail to the machine and tell it to send it to the desired person while driving the car simultaneously. Also Dictation Module can be proven very useful to visually impaired or physically debilitated people.

**Keywords:** Artificial Intelligence, Machine Learning, Voice Recognition, Speech Recognition, Speech To Text conversion

## 1. INTRODUCTION

The Earth is passing through a purplish patch of technology, where there is increasing demand of intelligence. In the 21st Century humans are surrounded by technology as they are the constituent of our day-to-day life cycle. In order to help this community in achieving multitasking, we are proposing Dictation Module system, which is basically the transcription of spoken text. It takes the dictation and converts it to the text document and send it as attachment to desired person as an email or send it to the printer for the purpose of printing and also acknowledges back in the form of voice. Dictation Module uses two basic technologies such as Speech recognition and Speech To Text. On the greater level or commercial level this can be achieved using a raspberry pi3 or some dedicated microprocessor.

## 2. LITERATURE SURVEY

1."Implementation of Text to Speech Conversion"

This paper includes implementation of Optical Character Recognition (OCR) in MATLAB. An image containing some characters is submitted to OCR and the text written into it is extracted to a text file. A feedforward neural network is used for classification resulting into recognition of text. The text is then converted to speech using Win 32 SAPI (Speech Application Programming Interface).

2."Design and implementation of Text to Speech for visually impaired people"

This paper talks about implementation of Text-To-Speech Synthesizer using Natural Language Processing (NLP) Module (which converts text into phonetic transcription along with prosody) and Digital Signal Processing (DSP) Module (which transforms symbolic information or phonetic transcription received from NLP into audible and intelligible speech).

3."VOICE RECOGNITION SYSTEM: SPEECH-TO-TEXT"

In this paper, for Speech-To-Text conversion they use Mel Frequency Cepstral Coefficients (MFCC) for feature extraction, Vector Quantization (VQ) for feature mapping, Hidden Markov Model (HMM) to create models for each letter and Viterbi Algorithm for feature testing of the dataset. To show desired text output MATLAB interface is used.

4."A Study of Speech Recognition"
Various Speech Recognition approaches (The acoustic-phonetic approach, Dynamic time warping (DTW), Neural Network based approach, etc.) are discussed in the given paper. Also some speech feature extraction techniques such as Linear Predictive Coding (LPC), Mel Frequency Cepstral Coefficients (MFCC), Signal Subspace Method are stated.



This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 1, January 2021

DOI 10.17148/IJARCCE.2021.10141

5.“Text to Speech Conversion”

Text-To-Speech Synthesizer is developed based on Raspberry Pi V2. The process involves image processing (where Tesseract OCR converts .jpg to .txt form) and voice processing (Festival software is used to convert the .txt to speech).

6.“Speech to text and text to speech recognition systems-A review”

This paper discusses various techniques for Speech Recognition (SR) and Speech-To-Text (STT) conversion as follows:

1) SR : Feature Extraction
- Linear Predictive Coding (LPC)
- Mel-Frequency Cestrum Co-efficient (MFCC)
- Dynamic Time Warping (DTW)

2) SR : Pattern Matching
- Template Based
- Knowledge Based
- Neural Based
- Statistical Based
- Hidden Markov Model (HMM)

3) STT
- Artificial Neural Network based Cuckoo Search Optimization

7.“Journal of Speech to Text Conversion”

This paper proposes a system ScribeBot which is a chatbot implemented using Raspberry Pi 3. It uses advanced deep learning algorithms and neural network (DNN) implemented in Google Speech API. It also uses Hidden Markov Model (HMM).

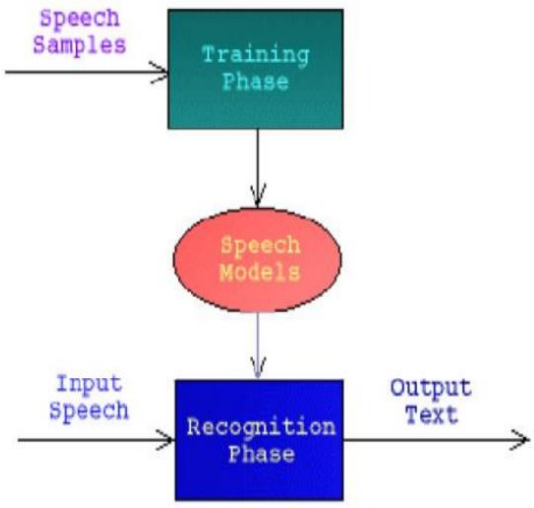
8.“A Smartphone-Based Multi-Functional Speech-To-Text Transcription System”

In this paper a smartphone-based speech-to-text system with LCD display is proposed. It works as follows –

1) The microphone of the smartphone inputs the speech and sends it to the android app.
2)The recorded speech is transcribed to English sounds using the android application with the support of Google voice engine which matches the recorded words to the stored words in its database.
3) Resulting text is then sent wirelessly to the Arduino-based hardware system via the Bluetooth.
4) Output text is then displayed on a 16x2 LCD module.

## 3. EXISTING SYSTEM

It is an online Speech to Text engine which uses Hidden Markov Model (HMM). It works in following phases.






This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 1, January 2021

DOI 10.17148/IJARCCE.2021.10141

**Phase 1 – Training phase**
System learns reference patterns which represent different speech sounds (e.g. phrases, words, phones) that constitute the vocabulary of the application.

**Phase 2 – Recognition phase**
Unknown input pattern is identified using set of references.
Speech Recognition System works in following stages -

- **Speech Analysis**

Speech data is analysed which includes speaker specific information due to vocal tract, excitation source and behaviour feature which is important for speaker recognition.

- **Feature Extraction**

Different individual characteristics of speech embedded in utterances are extracted.

- **Modelling**

Hidden Markov Model (HMM) is used to create models for each letter.

- **Testing**

Feature testing of the dataset is done.

The process for converting speech to text is done by using Raspberry pi and on the terminal of Raspbian image installed on raspberry pi. The disadvantage of this system is it can not run offlline.

## 4. PROPOSED SYSTEM

This project involves large set of data in the form of voices that has to be used to train the system, so that it should be able to recognize the voice of the user of system and differentiate between background noise and user's voice. The goal is to use this trained system to be a personal assistant to the user and user should be able to do some specific tasks with voice commands. The main task of this module is that, it should be able to understand what user said and convert it to a text format and write it into a text document which can be then made available for printing or sharing.
We are aiming this system should work without any dependencies of internet that means that it should be able to work independent of any network connection, all the systems, database as well as the program to differentiate the voices should be on the system local database. So, user can use it freely anywhere anytime and it will also reduce the cost of the project that will give the clients the financial edge the. Only dependency of such program or such system is that it will require a heavy configuration of hardware.

**The flow of the system is given as follows.**

- First of all, the system should be able to recognize the user's voice from all the background noises so that it can understand what he/ she is saying.
- The system should also be able to perfectly identify its user based on the voice and should not get confused with others' voices while recognizing user's voice as it will be questionable to the security of the system.
- The system will have the capacity to convert all the things said by user into text format or a text document which can be stored in the local database of the system to be used for printing or sharing purposes later on in the project.
- The system will have the capacity to email the particular file to anybody around the world using internet, as well as it should also be able to connect to the wireless or Internet Protocol printer around the world.
- There will be no user interface for this project as it will be voice controlled device, but by using the voice command the user will be able to login and log out from the system.
- We are choosing Linux platform to program this software and the languages used are Python, Bash.
- We choose these platforms as these have a large community support available on the internet and Python has some good libraries for us to achieve what is needed for the project.
- We prepare a dataset of speech samples from different speakers, with the speaker as a label , so it's a clear method of supervised learning. We add background noise to these samples to augment our data.
- We take the FFT(Fast Fourier Transform) of these samples.
- We train a 1D convent to predict the correct speaker given a noisy FFT speech samples.

All of these softwares will run on Tensorflow 2.3 or tf-nightly. The noise samples in the dataset need to be resampled to a sampling rate of 16000Hz, we can either accomplish this by using a software called FFmpeg or from the online audio converter. All the samples of the voice including the background noise are of 1 second long.



This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 1, January 2021

DOI 10.17148/IJARCCE.2021.10141

We are going to use PocketSphinx for Speech Recogntion and Speech To Text Conversion. It is a lightweight , open source continuous speech recognition engine. We are hoping for 90% accuracy for our model.

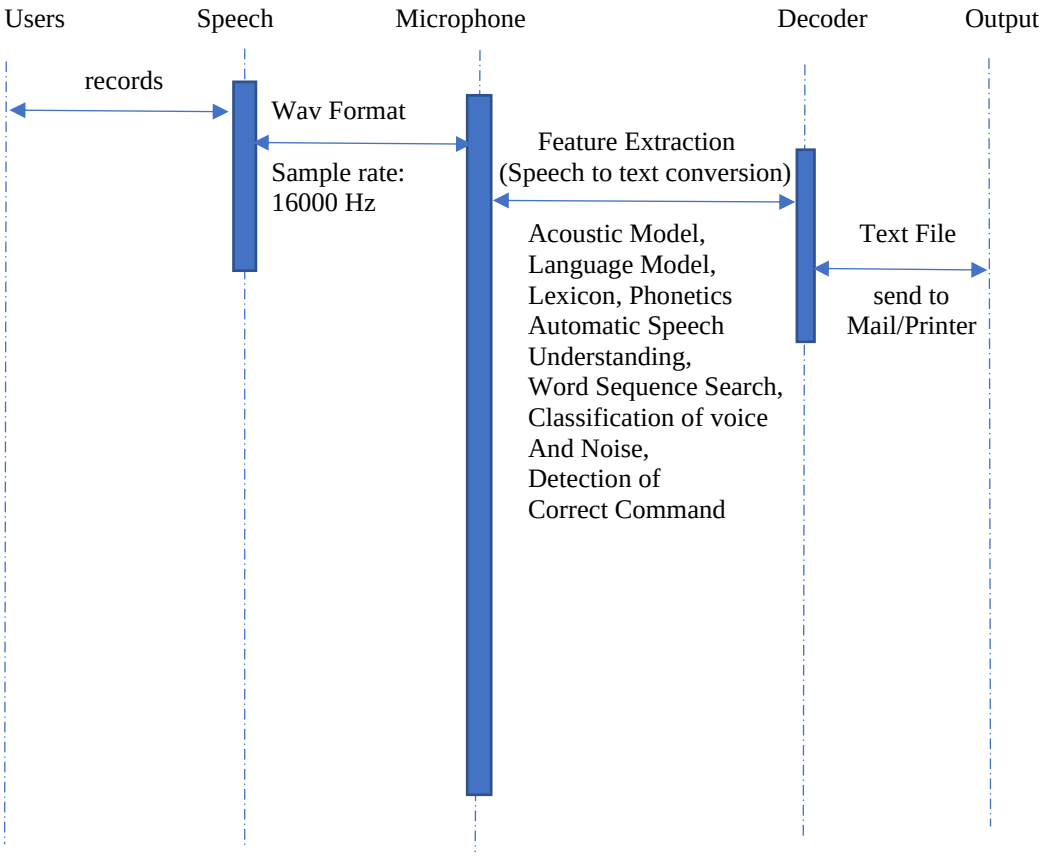



Fig: Sequence Diagram For Dictation Module

## 5. ML SYSTEM ARCHITECTURE

The System Architecture consists of the following,

- 1. Speech (by intended speaker after detection/recognition of speaker's voice)
- 2. Phonetic Dictionary
- 3. Acoustic Models
- 4. Feature Extraction
- 5. Decoding
- 6. Language Models
- 7. Word Sequence Search



This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 1, January 2021

DOI 10.17148/IJARCCE.2021.10141

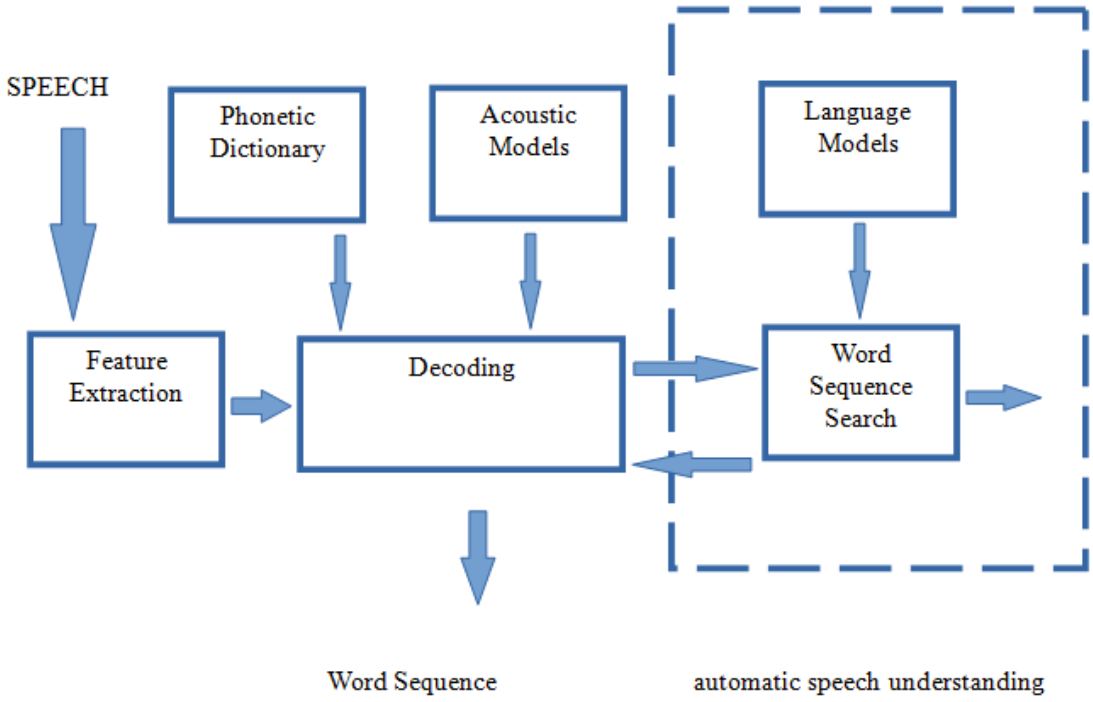



**Fig.1 ML System Architecture**

As shown in the above figure, Speech is recorded using microphone of the device and sent for Feature Extraction program in which we get the essential information from the speech and also properties like pitch and peaks in format of variable inputs values ranging from 0-255.
It is sent to decoding according to the phonetic dictionaries which are pre-saved in the memory of the device. It also depends upon the acoustic models created by the algorithms, also in addition we can send this decoded data or information to the automatic speech understanding program in which the program can auto-correct the grammar.
It can be online or offline, we prefer offline system because the dependency on internet gives us more disadvantages in the system.
We are also implementing text to speech, so that device can give out/speak out pre-coded sentences as to provide acknowledgement to the user.
The dictated text is stored in word document or simple text file and can be mailed or printed using voice command recognized by the system giving a kind of essence of IoT to the project.

## 6. CONCLUSION

This paper gives a descent approach for Speech To Text conversion after studying various researches done by multiple researchers in this field. We have given brief idea about how our system is going to work. Our Proposed System aims to be very useful for visually impaired people to interact with other people of the society as well as lawyers and businessmen for whom dictation technology is really necessary to speed up their work. It is also very user friendly and can be proven time saving method for intended users. This approach focuses only on offline processing which makes it useful on larger scale as there is no dependency on internet.
We look forward to do more research about this field and try and implement this system for more than one language and with whatever additional functionalities we could add.

## REFERENCES

[1] Chaw Su Thu Thu , Theingi Zin "Implementation of Text to Speech Conversion" International Journal of Engineering Research & Technology (IJERT) IJERT ISSN: 2278-0181 Vol. 3 Issue 3, March – 2014

[2] Itunuoluwa Isewon, Jelili Oyelade, Olufunke Oladipupo "Design and Implementation of Text To Speech Conversion for Visually Impaired People" International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 7– No. 2, April 2014



This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 1, January 2021

DOI 10.17148/IJARCCE.2021.10141

[3] Prerana Das, Kakali Acharjee, Pranab Das and Vijay Prasad "VOICE RECOGNITION SYSTEM: SPEECH-TO-TEXT" Journal of Applied and Fundamental Sciences

[4] Kaladharan N, "A study of Speech Recognition" International Journal of Innovative Research in Computer and Communication Engineering, Volume 3, Issue 9, September-2015

[5] S.Venkateswarlu, D.B.k. Kamesh, J.K.R Sastry, and Radhika Rani, "Text to speech conversion" Indian Journal of Science and Technology, Vol 9(38), DOI: 10.1.7485/ijst/2016/v9i38/102967, October 2016

[6] Ayushi Trivedi,Navya Pant, Pinal Shah,Simran Sonik and Supriya Agrawal "Speech to text and text to speech recognition systems-A review" IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 20, Issue 2, Ver. I (Mar.- Apr. 2018), PP 36-43

[7] Dhanush Kumar S, Lavanya S, Madhumita G, Mercy Rajaselvi V(2018): Journal of Speech to Text Conversion. International Journal of Advance Research, Ideas and Innovations in Technology. ISSN: 2454-132X Impact factor: 4.295 (Volume 4, Issue 2)

[8] Abayomi O. Agbeyangi & Adam B. Olorunlomerue (2019): A Smartphone-Based Multi-Functional Speech-To-Text Transcription System. Proceedings of the 15th iSTEAMS Research Nexus Conference, Chrisland University, Abeokuta, Nigeria, 16th – 18th April, 2019. Pp 165-174

[9] David Huggins-Daines, Mohit Kumar, Arthur Chan, Alan W Black, Mosur Ravishankar, and Alex I. Rudnicky, "POCKETSPHINX: A FREE, REAL-TIME CONTINUOUS SPEECH RECOGNITION SYSTEM FOR HAND-HELD DEVICES", Carnegie Mellon University Language Technologies Institute 5000 Forbes Avenue, Pittsburgh, PA, USA 15213, ICASSP 2006.



This work is licensed under a Creative Commons Attribution 4.0 International License