International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

A REVIEW ON METHODS FOR SPEECH-TO-TEXT AND TEXT-TO-SPEECH CONVERSION

Shivangi Nagdewani, Ashika Jain

[1]*UG Student, Computer Science Department, Inderprastha Engineering College, Ghaziabad, Uttar Pradesh, India.*
[2]*UG Student, Computer Science Department, Inderprastha Engineering College, Ghaziabad, Uttar Pradesh, India.*

Abstract - *Internet has evolved over time and has revolutionized many fields and impacted many lives. Internet is a boon to mankind. The main field revolutionized by the internet is communication. Internet has enabled faster and easier communication. Through this paper we aim to study the different methodology for Speech-To- Text and Text-To-Speech conversion that will be used in a voice-based email system. This system is based on interactive voice response. The aim is to study and compare the various methods used for STT and TTS conversions and to figure out the most efficient technique that can be adapted for both the conversion processes. As a result, based on review study it is found that HMM is a statistical model therefore most suitable for both STT and TTS conversions. At last a model using HMM and ANN methods for STT and HMM for TTS conversions proposed.*

Key Words: Speech to Text (STT), Text to Speech (TTS), Hidden Markov Model (HMM), Interactive Voice Response (IVR), Artificial Neural Network (ANN).

1. INTRODUCTION

The objective of the paper is to propose a layout for development of an interactive voice response-based mailing system that enables users to manage their email accounts using audio commands only. This paper will provide an analysis of various methods used for Speech-To-Text and Text-To Speech conversion.

In STT conversion the system detects words as well as phrases in audio input by a person or machine and converts them into a readable text format. This facilitates an efficient human to human, human to machine and machine to human communication. It is majorly useful when people of different languages and dialects communicate or interact with each other. In the absence of a STT conversion system people with different languages and dialects may not understand the words spoken by each other. Hence in such a scenario a STT convertor may help by detecting the words spoken by a person with different accent or dialect into text form which is easily readable and understood by the other person. So to achieve the above mentioned functionality many methods have been proposed and applied. Figure 1 represents basic steps of the STT process. Firstly important features are extracted from the input speech and then word and sentence matching is done using acoustic word models and defined syntax and semantic for the sentences. This process is mutually exclusive and can be done in parallel. At last the language modelling is performed using the selected modelling method.

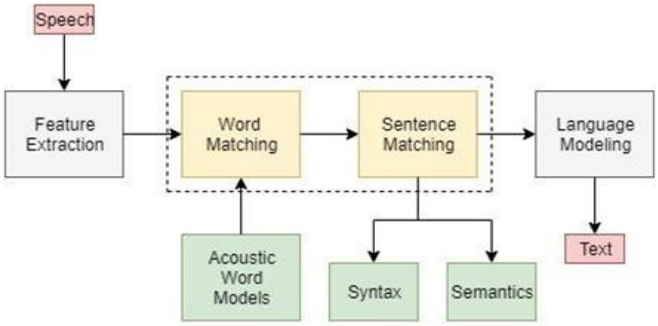

Fig.1.Speech To Text Process

In Text-To-Speech conversion the input text is analysed and then this text is converted into its audio version to play. This functionality has an effective advantage when a person understands a language but is not fluent with reading and writing in that language and is also useful for the people who are visually impaired as they cannot read but understand the message by hearing it. Figure 2 represents the basic steps of the TTS process. Firstly, the text is prepared for audio conversion by performing pre-processing and text normalization. To generate the Waveform of the text message the linguistic analysis and prosodic prediction is done in series.

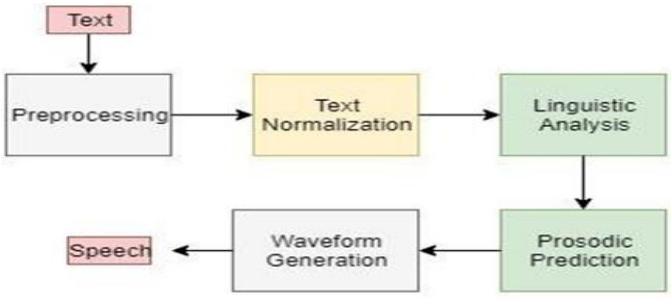

Fig.2.Text To Speech Process

2. MOTIVATION

A brief description of various research papers that were examined for this study is given below. The given below table 1 represents the summarization of various methods applied for Speech-To-Text and Text-To- Speech conversion. Going

through these papers it was observed that there is an additional scope of work on the STT and TTS conversion method.

Table -1: Summarization of various methods applied for Speech-To-Text and Text-To- Speech conversion

S. No.	Techniques Used	Description
1.	TTS,STT Conversions and IVR	[1] They suggested that for STT conversion the audio message should first be recorded and then be converted to text form and for TTS conversion the text should be translated to the audio and then play the audio message to the user. The proposed idea of an email system based on voice, makes use of 3 modules, namely: STT conversion, TTS conversion and IVR
2.	TTS,STT Conversions and IVR	[2]The proposed system focuses on providing user friendly platform to it users. The system implements Interactive-Voice-Response technology. In this system a pre-recorded voice will indicate the user to do some functions to avail some services.
3.	STT, TTS, Face recognition	[3]The paper proposes to develop a system that enables visually impaired, blind and people to use email facility as efficiently as some normal user. The dependency of the system on mouse or keyboard is almost diminished and it work on STT and TTS processes. Face Recognition is also used for authenticating the user identity.
4.	MFCC and HMM	[4] Proposed a STT system replacing traditional MFCC with HMM. The conventional MFCC approach was less efficient in extracting the features from the speech signals hence a new approach was suggested using HMM. The features passed to the HMM network resulted in better feature recognition from the input audio in contrast with the MFCC method. HMM exhibited vast improvement in the quality of feature extraction from the audio resulting in better computational time and accuracy for a Speech-To-Text conversion system.
5.	Automatic Speech Recognition, HMM model and human machine interface	The paper studied the deployment of STT by HMM and suggested to develop a machine interface system that depends on voice. The system could be deployed for helping 2 types of users: • People with disability who cannot access their email through use of mouse and keyboard, this category of users will be benefitted by the usage of a Speech-to-Text conversion system. • People who do not understand English or are not efficient in English and feel good to communicate in their native language i.e. English, Punjabi, Hindi.
6.	Pattern Recognition, Neural Network, Artificial intelligence	They suggested a number of speech representation and classification methods. A number of feature extraction techniques were also deployed by them along with database evaluation and performance. The analyzed the various concerns related to Automatic-Speech-Recognition and proposed methods to resolve them. The various methods to speech recognition addressed by them are: the AI Approach, the pattern recognition Approach and acoustic phonetic approach.
7.	ANN and HMM	[6] Suggested rate of STT conversion can be made better using various techniques together and better-quality of text can be obtained. The objective is to develop a continuous STT system that has a much wider vocabulary and is speaker independent that can detect voice of different speakers with precision. For developing such a system, a combination of ANN and HMM will be used highly.
8.	Speech Recognition, Feature Extraction, MFCC, Dynamic Time Wrapping (DTW)	[8] Detailed study shows that as the performance and reliability of the system are affected; some limitations are induced in the system. The study also illustrates that the maximum work for STT is carrier out for English language. Less work has been done for Indian and other regional languages. The study also examined that the rate of speech recognition is highest for English as compared to any other language. The reason for low recognition rate of Indian languages is due to their phonetic nature.
9.	Machine Learning, ANN ASR, Cuck Search Algorithm	[9] The paper summarizes the basic processes involved in a STT system which covers architecture of ASR(Automated Speech Recognition). The main focus for this paper is using Machine Learning in ASR, SVM, ANN with Cuckoo search algorithm along with ANN and

		back propagation classifier. The basic phases like: pre-processing, extraction of features and classification, of the STT system are studied by using machine learning. According to the generated results Hybridization of an algorithm with an optimization technique is considered better technique, traditional classifier results can be further improved by doing hybridization of it with other algorithms for optimization.
10.	Text processing, Text-To-Speech (TTS) synthesizer, Speech Enhancement	[10] Suggested that Test-to-Speech synthesizer is developing rapidly from past few years to gain the current shape. The most suitable methods for TTS are Formant, Articulator and Concatenative synthesis. Even in India some research organizations are also working on Text-to-Speech in regional languages like Marathi, Hindi, Telugu, Punjabi, Kannada, so on. A vast scope of improvement can be achieved in TSS synthesis to obtain a good amount of natural and emotion aspect.
11.	TTS	[12] Suggest the usage of TTS synthesis for English language. A speech synthesizer is required for converting the text data to speech. A TTS system has 2 important processes: First, handling of text data and seconds is speech generation. The successful and satisfactory results are obtained by .net framework system. The system can be deployed in the email readings, web applications, mobile applications.
12.	Text-to-Speech conversion (TTS), Speech synthesis, Syllabification, Concatenation, Text Normalization, Text Conversion	[13] The paper aims to create a TTS system for native languages like Hindi. The system involves of 2 main steps: Text Pre-Processing and Speech Generation. A Concatenative synthesis-based approach is considered for obtaining the speech from the text. A spellchecker module is also implemented for checking the correctness of words for native languages like Hindi.
13.	HMM, ANN, DWT	[11] They Examined various techniques for TTS and STT. After examining various STT synthesis, TTS synthesis and speech translation systems they concluded that: • In STT, HMM works as a better generator of text from speech despite its drawbacks due their computational feasibility. • In TTS systems formant synthesis that makes use parallel and cascade synthesis works as the best converter. The wide usage of Hybrid machine translation is due to its inculcation of advantages of both statistical and rule-based machine techniques for translation. The system takes care that it creates text which syntactically and grammatically correct also taking care of the smoothness in a text, fast learning ability and data acquisition.

By Analyzing the various papers, we have concluded that there is vast scope of evolution in the domain of Text- to-speech and Speech-to-text conversion. In the next section we have analyzed various TTS and TTS synthesis methods.

3. COMPARISION BETWEEN VARIOUS MODELS

Table -2: The various models for Speech-To-Text Conversion

METHOD	ADVANTAGE	DISADVANTAGE
Linear Predictive Coding (LPC	• LPC is a Static approach used for feature extraction. • The concept of LPC is that it can take the voice sample as linear combination combining past voice samples. • The voice signal is fragmented into N frames and then these framed windows are converted into text.	• Uses fixed resolution spectral analysis along with a subjective frequency scale.
Mel-Frequency Cestrum Co- efficient (MFCC)	• MFCC is another approach based on extracting features of signal by using filter bank. • The technique applies steps like Framing, Windowing and Discrete Fourier Transform for STT conversion.	• The problem with MFCC that it requires Normalization as values in MFCC are not very efficient in existence of surroundings or additive noises.

Dynamic Time Wrapping	• The DTW algorithm is used to find the analogy in two-time series events that vary in speed by using dynamic programming. • Its purpose is to iterate the pair of sequence of feature vectors and finding a feasible match between them.	• The problem arises in selecting the reference template for comparing the time series events.
Hidden Markov Model	• HMM is a statistical model used for STT conversion. • HMM exhibits its own structure and self -learning which makes them very useful for STT conversion.	• In this method, the voice signal is seen as a static signal or short-term time static signal. • HMM is serial.
Neural Network	• Neural network is also a statistical model, represented as a graph. • Neural networks make use of connection functions values and connection strengths for the state transactions.	• Here in neural network model, ANN are parallel.
Hybrid Approach	• The proposed hybrid approach is used for Speech to Text conversion because speech frequencies are in parallel, whereas syllable series and words are in serial. • This shows that both the methods are useful indifferent context. Both HMM and Neural Networks techniques are implemented together. • As Neural networks show good performance in studying the probability from parallel voice input and Markov models can use the phoneme observation probabilities that neural networks provide to produce the possible phoneme sequence or word.	

Table -3: The various approaches for Text-To- Speech conversion

METHOD	ADVANTAGE	DISADVANTAGE
Rule Based Machine Translation (RBMT)	• RBMT makes use of syntactic and semantic analysis for conversion of text to speech. • The system is collection of Grammatical rules. • It performs a lookup of each word present in the input text consisting the Grammar and Dictionary base of the particular language to perform a TTS conversion.	• The RBMT is inefficient for big systems.
Statistical Machine Translation (SMT)	• SMT is probabilistic technique using Bayes Theorem that assigns each sentence in the input with a probability. • The more the value of probability the more is the efficiency in conversion of that sentence into the speech format.	• The disadvantage of this approach is the high cost involvement and it doesn't work well enough for different languages.
Hidden Markov Model (HMM	• HMM is a probabilistic technique similar to SMT but given better accuracy for TTS conversion. • HMM can be deployed for both voice recognition systems and also text-to-speech synthesis systems to generate an audio signal from text input. • The advantage of adopting HMM is, it is an automatically trained network.	

By analyzing the various methods for STT and TTS we have examined that HMM provides maximum efficiency for STT and TTS conversion. Also a optimal amount of efficiency is provided by neural network for STT. Hence we have proposed the Hybrid approach for STT conversion that deploys HMM and Neural network and for TTS, the HMM model provides the highest accuracy in comparison to others.

4. THE HIDDEN MARKOV MODEL

HMM are a sub class of the dynamic Bayesian models and exhibit their own kind of structure which makes them very useful for a large number of applications. The system that being modelled using the HMM has a Markov process having some hidden states and hence named as HMM. An HMM model in the most basic form can be assumed as a probabilistic model some state variable S and some observation variable O and transition between the states, each transition having an associated probability. In simple words, HMM belong to graphical model category that are efficient in predicting some hidden variables from some of visible variables. A basic example to quote on HMM is weather prediction for a location on the basis of the clothes that the people living there are wearing. The main advantage of using the HMM is Markov assumption that states "The future event is completely independent from the past event and depends only the present". In simple words if we are aware of the current state, we need not have the training data to predict our future state. The HMM are mostly deployed in reinforcement learning projects like Speech Recognition, Text Recognition, Robot Localization, Biological sequence analysis, Handwriting Recognition, Pattern Recognition, etc. STT and TTS are particularly HMM based problem.

5. PROPOSED MODEL

The model aims to develop a system that provides the user with 2 functionalities: Firstly, to send an Email by converting the voice input message from the user into text and sending it. Secondly, to converting the Text at the recipients end to voice output and narrating the message to user.

The model for STT conversion is carried out by HMM and Neural Network as it gives the highest accuracy for STT.

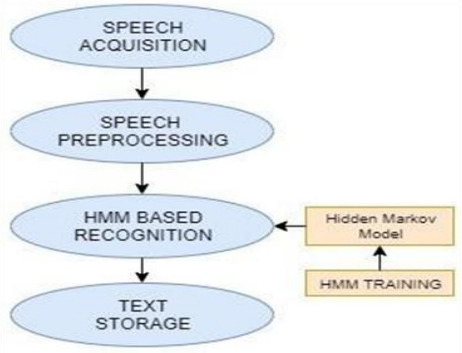

Fig.3.STT using HMM

Figure 3 represents basic steps for STT conversion using HMM. It consists of 4 steps:

1. Speech Acquisition: Getting the Speech data as input. The speech is taken as input using the microphone and is stored in the memory.
2. Speech preprocessing: The noisy environment and disturbances in speech signal are removed and it results in actual speech for conversion. By voice activity detection, these pauses are removed from the voice input.
3. HMM based recognition: The HMM is built by the system for each word in the vocabulary and these models are trained at the time of training phase. From speech preprocessing to HMM model construction, the steps of training are performed and generated HMM is loaded.
4. Text Storage: The matched text is stored as output and assembled to generate the text output.
The model for TTS conversion is also implemented through HMM training for effective results.

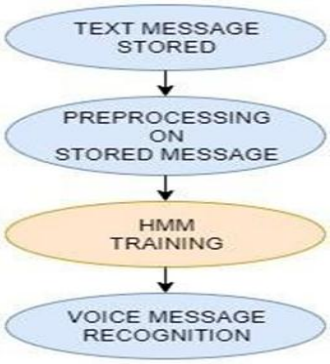

Fig.4.TTS Using HMM

Figure 4 represents the basic steps for TTS conversion using HMM. It consists of 4 steps:

1.Text Message Storage: The text is taken as input at runtime from the user and stored into the system.
2. Pre-processing of Stored message: This step is useful for removing the irregularities and noisy data. Also, the text input splits into the different overlapping text frames.
3. HMM training: HMM is used for text recognition and is a statistical model for modelling an unknown system using an

observed output sequence. The HMM training involves pattern recognition for mapping the words to their sounds and creation of pattern representation of the features extracted from text class using one or more test patterns that correlates to speech sound of the same class.

4.Voice Message Recognition: Finally, the speech is generated as output and assembled which results in the voice output.

The reason for adapting HMM for both the TTS and STT processes is that it is a statistical model and provides a simple and effective framework for modeling temporal vector sequences. HMM provides a natural framework for building such models [14]. HMM depends on currently executing events and does not depends on the previous events. Also, the HMM is a really Dynamic Model which has training and self- adapting capabilities that increases the quality to STT and TTS. The HMM lies in the heart of all modern STT and TTS recognition systems.

5. CONCLUSION

Various techniques exist for STT and TTS synthesis. By deploying the Hidden Markov Model technique, the rate of STT and TTS processes can be improved and better-quality speech and text must get generated. The most suitable technique for STT conversion is by deploying a combination of Hidden Markov Model with Deep Neural Network, which can be implemented in Python using Google's Speech Recognition API module. This system can be improved by considering the punctuation marks while converting speech to text. The best method for TTS conversion is by deploying the HMM model that gives the best accuracy and can be implemented in Python using pyttsx3 or gTTS modules. This system of Text-To-Speech and Speech-To-Text can be implemented for to different languages like English, Hindi, Punjabi etc, depending upon the user's requirement and has the capability to recognize these languages and change it to the desired text or speech format.

6. FUTURE SCOPE

The By evaluating the various methods for TTS and STT, we have concluded that HMM works well for both STT and TTS. Using the STT and TTS by HMM, a web-based application can be created for sending and viewing voice-based messages. In a voice-based email system user must be allowed to send and receive emails on the go by only using their voice. The application will be designed in a manner that it will instruct the user with voice instructions to do some functions and the user will react accordingly. Besides using IVR to provide the instructions to the user the system also uses the latest technology of Speech-To-Text and Text-To-Speech conversion. The advantage of this application is, the manual labor of typing is completely reduced to absolute 0 and the user will only have to respond through voice-based commands. So, the Voice Based E-mail System using STT and TTS will cut the manual labor of typing at the part of the user. This system can be deployed for the purpose of effective communication by normal, illiterate and visually challenged people.

ACKNOWLEDGEMENT

This review on methods for Speech to Text and Text to Speech conversion was supported by Ms. Shelly Gupta, Assistant Professor, Inderprastha Engineering College, Ghaziabad, Uttar Pradesh, India. We thank Ms. Shelly Gupta for providing insight and expertise that greatly assisted this review paper.

REFERENCES

1. H. K. ,. A. L. Pranjal Ingle, "Voice based e-mail system for the Blinds".
2. A. ,. A. a. K. T.Shabana, "A Review on Voice based e-mail System for Blinds –," 2015.
3. N. K. P. S. Shashank Tripathi, "Voice based Email System for Visually Impaired and Differently abled," International Journal of Engineering Research & Technology (IJERT), 2019.
4. D. Y. S. R. Ibrahim Patel, "SPEECH RECOGNITION USING HMM WITH MFCC- AN," 2010.
5. N. K. P. K. Bhupinder Singh, "Speech Recognition with Hidden Markov Model: A Review," International Journal of Advanced Research in Computer Science and Software Engineering, 2012.
6. S. K. Saksamudre, "A Review on Different Approaches for Speech," 2015.
7. G. K. K. Sanjivani S. Bhabad, "An Overview of Technical Progress in Speech Recognition," International Journal of Advanced Research in Computer Science and Software Engineering, 2013.
8. P. K. Kurzekar, "A Comparative Study of Feature Extraction Techniques for Speech Recognition System," International Journal of Innovative Research in Science, Engineering and Technology, 2014.
9. N. T. D. B. Sunanda Mendiratta, "A Robust Isolated Automatic Speech Recognition System using Machine Learning Techniques," 2019.
10. S. R. Mache, "Review on Text-To-Speech Synthesizer," International Journal of Advanced Research in Computer and Communication Engineering, 2015.
11. 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), 2018.
12. K. N, "An English Text to Speech Conversion System," 2015.
13. Kaveri Kamble, Ramesh Kagalkar, "A Review: Translation of Text to Speech Conversation for Hindi Language," International Journal of Science and Research (IJSR) , 2012.
14. Mark Gales and Steve Young, The Application of Hidden Markov Models in Speech Recognition, Foundations and Trends in Signal Processing, 2008.