# A Review on Text-to-Speech Converter

Dr. S.A. Ubale[1], Girish Bhosale[2], Ganesh Nehe[3], Avinash Hubale[4], Avdhoot Walunjkar[5]
[1,2,3,4,5] *AI&DS Department, Zeal College of Engineering & Research, Pune, India*

***Abstract*— The Internet is a bone to mankind. The main field revolutionised by the internet is communication. A Text-to-speech synthesiser is used to convert text into speech (voice) by analysing and processing the text using Natural Language Processing and then using Digital Signal Processing technology to convert this processed text into a synthesised speech representation of the text. Through this paper, we aim to study the different methodologies for Speech- To-Text and Text-To-Speech conversion that will be used in a voice-based email system. Developed a useful Text-to- Speech synthesiser in the form of a simple application that converts inputted text into synthesised speech and reads it out to the user, which can then be saved as an mp3. file. The development of a text-to-speech synthesiser will be of great help to people with visual blindness and make reading through large volumes of text easier.**

***Index Terms:*** **Text to Speech, Python, Audio.**

## I.INTRODUCTION

Text-to-speech (TTS) technology reads aloud digital text. It can take words on computers, smartphones, tablets and convert them into audio. Converting Text to Speech basically refers to a program where you give input as a text and the output you receive is the input text in the form of a speech. It is a Python library to interface with text to speech using different modules in python programming language. It does not require an Internet connection and it's pretty easy to use. Through this paper, the aim is to study the different methodologies for Speech-To-Text and Text-To-Speech conversion that will be used in a voice-based email system.

## II.LITERATURE REVIEW

In [1] S. R. Mache 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. In [2] N. K. P. S. Shashank Tripathi proposes 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. 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 analysed 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. In [2] N. K. P. S. Shashank Tripathi proposes that systems are trained by the individual who will be using the system. These systems are capable of achieving a high command count and better than 95% accuracy for word recognition. The drawback to this approach is that the system only responds accurately only to the individual who trained the system. This is the most common approach employed in software for personal computers. In [3] Rubin, P., Baer, T., and Mermelstein, P proposes that the synthesized speech can be created by concatenating pieces of recorded speech that are stored in a database. Systems differ in the size of the stored speech units; a system that stores phones or diphones provides the largest output range, but may lack clarity. For specific usage domains, the storage of entire words or sentences allows for high- quality output. Alternatively, a synthesizer can incorporate a model of the vocal tract and other human voice characteristics to create a completely "synthetic" voice output. In [4] Van Santen, J.P.H., Sproat, Olive, J.P., and Hirschberg

suggested that the quality of a speech synthesizer is judged by its similarity to the human voice and by its ability to be understood. An intelligible text-to-speech program allows people with visual impairments or reading disabilities to listen to written works on a home computer. A text-to-speech system (or "engine") is composed of two parts: a front-end and a back-end. In [5] Mattingly I. G. proposes that the system can be deployed in the email readings, web applications, mobile applications. Generating synthetic speech has been a curiosity for the past 100 years. Around those years Gerbert of Aurillac created the first known mechanical talking machine. For the next two centuries, inventors like Albertus Magnus and Roger Bacon created machines known as “talking heads”. In [6] Kaveri Kamble, Ramesh Kagalkar suggested 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. By analysing 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 analysed various TTS and TTS synthesis methods.

## III.SYSTEM DESIGN

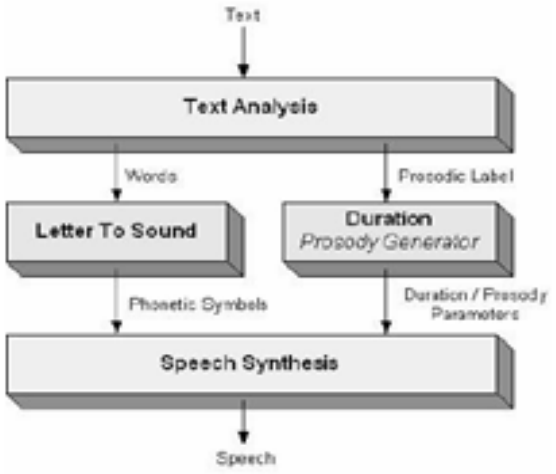


Fig 1.1 System architecture

A speech synthesis system is by definition a system, which produces synthetic speech. It is implicitly clear, that this involves some sort of input. What is not clear is the type of this input. If the input is plain text, which does not contain additional phonetic and/or phonological information the system may be called a text-to-speech (TTS) system. The TTS system converts an arbitrary ASCII text to speech. The first step involves extracting the phonetic components of the message, and we obtain a string of symbols representing sound-units (phonemes or allophones), boundaries between words, phrases and sentences along with a set of prosody markers (indicating the speed, the intonation etc.). The second step consists of finding the match between the sequence of symbols and appropriate items stored in the phonetic inventory and binding them together to form the acoustic signal for the voice output device. A schematic of the text-to-speech process is shown in the figure 1.1.

## IV.METHODOLOGY

For building the Text-to-Speech converter pyttsx3 library is used which is platform independent. The major advantage of using this library for text-to-speech conversion is that it works offline. However, pyttsx supports only Python 2.x. Hence, we will see pyttsx3 which is modified to work on both Python 2.x and Python 3.x.

1 Speech synthesis techniques will be used in order to get the naturalness quality in the synthetic speech.
2 The process of the English language can be Used as the basic unit for speech synthesis.
3 Speech database for the English language will be developed using phoneme.
4 Phonemes will be searched in the database and corresponding phonemes sounds will be Concatenated to generate synthesized output Speech.

Advantages of this implementation:

a. Text-to-Speech converter minimizes human agent workload, provides personalized services, accelerates throughput, and reduces operational costs.
b. Personalize the pitch of your selected voice, up to 20 semitones more or less from the default. Adjust your speaking rate to be 4x faster or slower than the normal rate.
c. Extend the reach of your content – TTS gives access to your content to a greater population,

such as those with literacy difficulties, learning disabilities, reduced vision and those learning a language. It also opens doors to anyone else looking for easier ways to access digital content.

Disadvantages of this implementation:

a. The resulting speech is less than natural and emotionless. This is because it is impossible to get audio recordings of all possible words spoken in all the possible combinations of emotions, prosody, stress etc.
b. Pronunciation analysis from written text is a major concern.

## V.CONCLUSION

A speech-to-text conversion is a useful tool that is on its way to becoming commonplace. With Python, one of the most popular programming languages in the world, it's easy to create applications with this tool. As we make progress in this area, we're laying the groundwork for a future in which digital information may be accessed not just with a fingertip but also with a spoken command.

## VI.FUTURE SCOPE

The existing systems encounters issues while performing scan on documents with complex backgrounds and the output is expected to have less accuracy. The proposed system ensures to read text present in the image for assisting blind people. Pre-processing part ensures efficient background separation with an improved algorithm. The future work will be concentrated on developing an efficient product that can convert the text in the image to speech with high accuracy.

## REFERENCES

[1] S. R. Mache, "Review on Text-To-Speech Synthesizer," International Journal of Advanced Research in Computer and Communication Engineering, 2015.
[2] N. K. P. S. Shashank Tripathi, "Voice based Email System for Visually Impaired and Differently abled," International Journal of Engineering Research & Technology (IJERT), 2019.
[3] Rubin, P., Baer, T., and Mermelstein, P., 1981. An articulatory synthesizer for perceptual research. Journal of the Acoustical Society of America 70: 321–328.
[4] Van Santen, J.P.H., Sproat, Olive, J.P., and Hirschberg, J., 1997. Progress in Speech Synthesis. Springer.
[5] Mattingly I. G., Speech Synthesis for Phonetic and Phonological Models, T.A. Sebeok (Ed.) Current Trends in Linguistics, Vol. 12, (1974) p. 2451-2487.
[6] Kaveri Kamble, Ramesh Kagalkar, "A Review: Translation of Text to Speech Conversation for Hindi Language," International Journal of Science and Research (IJSR), 2012.