A Survey on different approaches for Speech to Text and Text to Speech in Email System for Visually Impaired People

Sanskriti Naik
Computer Engineering Department
Shree Rayeshwar Institute of Engineering and Information Technology
Ponda, India

Nikhil Naik
Computer Engineering Department
Shree Rayeshwar Institute of Engineering and Information Technology
Ponda, India

Gopura Prabhu
Computer Engineering Department
Shree Rayeshwar Institute of Engineering and Information Technology
Ponda, India

Akash Bhayje
Computer Engineering Department
Shree Rayeshwar Institute of Engineering and Information Technology
Ponda, India

Vijaykumar Naik Pawar
Computer Engineering Department
Shree Rayeshwar Institute of Engineering and Information Technology
Ponda, India

Shailendra Aswale
Computer Engineering Department
Shree Rayeshwar Institute of Engineering and Information Technology
Ponda, India

ABSTRACT
Currently in the industry, communication is the most important element to move towards progress. As the world is moving fast towards digitization, so are the means of communication. Emails, video calls, phone calls, texts messages, social media etc. are an integral part in this tech-savvy world for conveying messages. Unfortunately, visually challenged people cannot make use of these technologies; hence this project is made for them. Voice based email apps make use of technologies such as Speech to Text and Text to Speech. Most of these apps use functions such as acoustic-based speech recognition, conversion from text signals to speech, and from speech to text signals, language translation amongst various others. In this review paper, we will be observing various techniques and algorithms that are applied to achieve the mentioned functionalities.

Keywords
Speech toText, Text to speech, Speech recognition, communication, Machine translation

1. INTRODUCTION
Voice is the common, efficient and basic method of communication for individuals to interact with others. Today, speech technologies are available for limited but interesting range of task. These are the technologies that enable machines to respond reliably and correctly to human's voice and provide very useful and much valuable services. As communication with computer is very fast using voice rather than using keyboard, so people prefer such system. Communication is being dominated by human beings by spoken language, therefore people will expect voice interfaces with computer.

Social media, science and email use technology. But not everyone is lucky to use these facilities because of various reasons. Illiterate and Visually impaired people face this problem a lot as they cannot see the screen, use keyboard or anything related to computers. They are dependent on other people for using the mail facilities like for composing mail, viewing the received emails, deleting mails etc. The existing systems are working on keyboard actions, mouse clicks, screen readers and voice commands. They are not fully voice command based but partially work on voice commands like recording voice while sending messages. Mouse is one of the key hardware devices that is involved in this kind of systems.

Screen readers were also used as they enabled visually impaired people to read their messages on the screen via Braille language. But this could not be implemented by the developers completely. People also had to rely on keyboard or mouse clicks for operating the system which could increase the complications for their use.

So, there must be some technique that would not cost much and still, be completely user friendly. Here the user is a blind person, so the criteria of user friendliness would be fulfilled if they are able to use this email system completely on their own. This could really help a lot of blind people as this system would help them to become independent in case of using technologies like this.

2. LITERATURE REVIEW
Speech is greatest and prime mode of message transmission amid human beings. Speech has the possibility of being vital method of interaction with machine. This paper gives an outline of major technical viewpoint in STT (Speech To Text) and TTS (Text To Speech) algorithm. This paper helps in selecting the method along with its comparative advantages and disadvantages.

3. SPEECH RECOGNITION
Speech Recognition is the capability of program to recognize words and phrases in verbal language and translate them into machine-readable format. Speech Recognition Systems can be classified on basis of the following parameters:

3.1 Basic Speech Recognition Model
i) Pre-processing: The analogue voice signal is transformed

into digital signals which is used for processing in future. The digital signal is stimulated to the first order filters to spectrally compress the signals. This benefits in growing the signal's energy at a higher frequency.

ii) Feature Extraction: This stage searches for the set of parameters of sounds that have an association with speech signals. These parameters, that are known as features, are figured by processing acoustic waveform. The crucial concentration is to calculate an arrangement of feature vectors provided that a compact illustration of the specified input signal. [4] Generally used feature extraction methods are given below:

• Linear Predictive Coding (LPC): The elementary idea is that the speech model can be approached as a linear blend of previous speech samples. The digitalized signal is jammed into frames of N samples. Then individual sample frame is windowed to minimalize signal cutoffs. Each of the framed windows are then auto-associated.[2]

• Mel-Frequency Cestrum Co-efficient (MFCC): It's a very influential procedure and practices human acoustic perception arrangement. MFCC applies convinced phases to the input signal: Framing: Speech wave- form is cropped to eliminate intrusion if existing; Windowing: curtails the breaks in the signal; Discrete Fourier Transform: alters individual frame from time domain to frequency domain; Mel Filter Bank Algorithm, the signal is plotted contrary to the Mel spectrum to impersonate human hearing.[2][4]

• Dynamic Time Warping: This process is used for measurement of the likeness among two-time sequences which might differ in speed, built on dynamic programming. It's purpose is to align two arrangements of feature vectors iteratively until an optimal match amongst them is found. [1]

iii) Acoustic Models: It is the essential part of Automated Speech Recognition (ASR) system where a linking between the audio information and phonetics is recognized.[4]Training establishes a link between the rudimentary speech elements and the auditory annotations.

iv) Language Models: This system induces the possibility of a word occurrence next to a word arrangement. It comprises the physical limitations available in the language to produce the possibilities of occurrence. The language model differentiates words and phrases that sound alike.[1][4]

v) Pattern Classification: It is the method of comparing the unidentified pattern with current sound reference pattern and calculating resemblance among them. After finishing the training of the system at the period of testing, patterns are classified to identify the speech. [4]Various approaches for pattern matching are:

• Knowledge Based Approach: This method takes set of features from the speech and then train the system to produce set of production rules from the samples. [1]

• Neural Network Based Approach: This method is skilled of resolving more complex recognition task. The elementary idea is to accumulate and in- corporate information from a variability of data bases with the problem.

• Statistical Based Approach: In this method, dissimilarities in speech are demonstrated statistically by means of training procedures.[1][3]

3.2 Speech to Text Conversion Methods

Speech to text conversion is the method of translating spoken words into texts. It is identical to speech recognition but the STT is used define the extensive procedure of speech understanding. STT follows similar ideologies and steps of speech recognition, with different mixtures of numerous procedures for each phase. Some broadly used alteration procedures are given below. [1][3]

i) Hidden Markov Model (HMM): HMM is a statistical model used in speech recognition since a speech signal can be seen as a section wise fixed signal or a short-time motionless signal. [1]HMM, models are suitable for real-time speech to text conversion for portable users. It depends on the following parameters:

• Recognition accuracy- Recognition is the method of comparison of the unknown assessment pattern with each sound class orientation pattern and figuring a measure of resemblance between the assessment pattern and each orientation pattern.

• HMM Recognition: It is the method of associating the unknown assessment pattern with each sound class reference pattern and calculating a measure of similarity. Extreme probability is used for recognition.[2]

ii) Artificial Neural Network: ASR is constructed for improved interface of human and computer interaction. For this, a following process is followed:

• Pre- processing of the speech signals is the crucial part of speech recognition which is implemented to eliminate unnecessary waveform of the signal. The signals are fed to the high-pass filters to get rid of the circumstantial noises.[1]

• Two types of acoustic features are mined, from the speech signal. They are Mel Frequency Cep- strum Coefficients (MFCC) and Linear Predictive Coding coefficients (LPCC).[1][2]

Table I. Comparison of speech recognition techniques

MODELS	TECHNIQUES	FINDINGS	ISSUES
SPEECH RECOGNITION: FEATURE EXTRACTION	Linear Predictive Coding (LPC)	Static feature extraction method. Spectral analysis is done with a fixed resolution along a subjective frequency scale.	Frequencies are weighted equally on a linear scale while the frequency sensitivity of the human ear is close to the logarithmic
	Mel-Frequency Cestrum Co-efficient (MFCC)	It is the nearest feature extraction method to the actual human auditory speech perception.	MFCC values are not very robust in the presence of additive noises. Normalization is required [1]
	Dynamic Time Warping (DTW)	It is used to cope with different speaking speed. Simple hardware implementation.	Difficulty in selecting the reference template.
	Knowledge Based	Uses the information regarding linguistic, phonetic and spectrogram	Explicit modelling variation in speech is difficult to obtain and use successfully, so, this approach is impractical
	Statistical based	Present models use this approach	Low accuracy of priori modelling presumption reducing its trend

4. TEXT TO SPEECH CONVERSATION

Text-To-Speech (or TTS) will deploy a string of text into an auditory clip. Text-To-Speech is a procedure in which input text is first evaluated, then handled and understood, and then the text is transformed to digitalized audio and then spoken.

• Text Processing: The input text is examined, standardized and transliterated into phonetic or language representation.

• Speech Synthesis:

Some of the speech synthesis techniques are voice rendering:

i) Articulator Synthesis:

Uses automatic and auditory model for speech generation. It generates comprehensible synthetic speech but it is far-off from expected sound and hence not extensively used.[1]

ii) Formant Synthesis:

In this model, illustration of discrete speech fragment is kept on a parametric foundation. There are two elementary structures in formant synthesis, parallel and cascade, but for improved performance, some kind of combination of these 2 structures is used. A cascade formant synthesizer consists of band-pass resonant circuit linked in series. The output of each formant resonant circuit is applied to the input of the following one.[1][4]

The cascade structure requires only formant occurrences as control info. A parallel formant synthesist comprises of resonant circuit connected in parallel.

iii) Concatenative Synthesis:

This method produces sound by concatenating brief samples of sound called units. It is used in speech synthesis to produce user specific arrangement of sound from a catalog built from the recording of other sequences.[1]

Table II. Comparison of text to speech conversion techniques

MODELS	TECHNIQUES	FINDINGS	ISSUES
TEXT TO SPEECH	Articulator synthesis	Use mechanical and acoustic model	Output is far from natural voice.
	Formant Synthesis	Based on the source filter- model of speech	The cascade structureshas been found better fornon-nasal voiced soundsand because it needs lesscontrol information thanparallel structure, it isthen simpler toimplement.
	Concatenative Synthesis	Duration of units isnot fixed, can bevaried as perimplementation.	Complex Method

6. PROPOSED METHODOLOGY

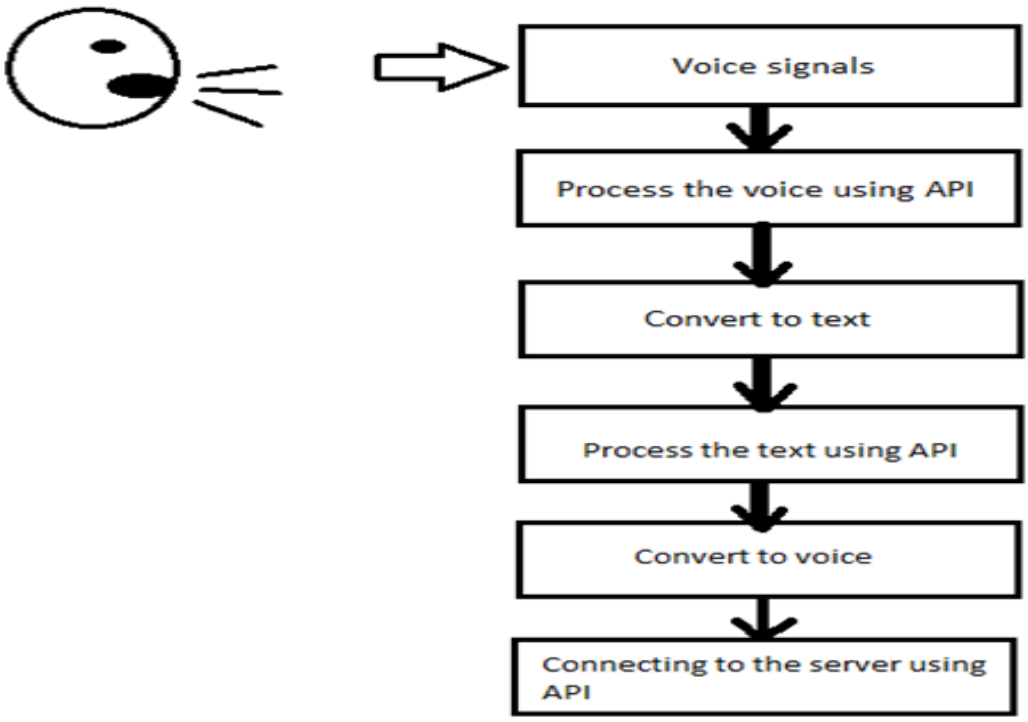

Fig 1: Overview of the system

This system is a voice-based email application by which a visually impaired person can easily transfer email through his/her voice. To create the application, numerous processing procedures will be employed to transfer the Voice Signals to mail server. They are:

i) Process the Voice using API
ii) Convert to Text
iii) Process the text using API
iv) Convert to Voice
v) Connecting to the server using Gmail API.

• Google Speech API- Audio is translated into text by applying powerful neural network models. This API identifies audio uploaded in the system and integrate with user's audio storage on Google Cloud Storage.

• Google Text to Speech API- Text is taken and a request is sent to Google Translate API Web servers which returns an audio file with text converted into speech.

• Gmail API: Gmail API gives the advantage of send mail and accessing Gmail mailbox. Highest importance is given to Gmail API by programmers for official access to a user's Gmail data.

7. ACKNOWLEDGEMENT

The authors would like to thank Prof.Vijaykumar Naik Pawar and the principal Prof.ShailendraAswale for their constant support and motivating the authors to complete this project.

8. CONCLUSION

This paper gives a clear idea about Speech-To-Text and Text-To-Speech. In Speech to text, HMM works as the best speech signal to text converter despite of its drawbacks because of their computational feasibility. Similarly, in Text-To-Speech, formant synthesis that makes use of cascade and parallel synthesis works as the best converter. This paper presents the development of existing speech to text and text to speech system by adding Google speech API, Google text to speech API and Gmail API to make a voice-based email system for visually impaired people. This framework of voice-based email system can be further developed to make this system completely voice based with no mouse clicks.

9. REFERENCES

[1] Ayushi Trivedi,Navya Pant, Pinal Shah,SimranSonik and Supriya Agrawal,Speech to text and text to speech recognition systems-Areview.

[2] R. Thiruvengatanadhan, Speech Recognition using SVM

[3] Prachi Khilari, Prof. Bhope V. P,a review on speech to text conversion methods.

[4] AnchalKatyal, Amanpreet Kaur, Jasmeen Gill, Automatic Speech Recognition: A Review.

[5] Ayushi Trivedi,Navya Pant, Pinal Shah,SimranSonik and Supriya Agrawal,Speech to text and text to speech recognition systems-Areview.

[6] R. Thiruvengatanadhan, Speech Recognition using SVM

[7] Prachi Khilari, Prof. Bhope V. P,a review on speech to text conversion methods.

[8] AnchalKatyal, Amanpreet Kaur, Jasmeen Gill, Automatic Speech Recognition: A Review.

[9] Suman K. Saksamudre, P.P. Shrishrimal, R.R. Deshmukh, A Review on Different Approaches for Speech Recognition System.