IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

IJARSCT

Impact Factor: 7.301

Volume 3, Issue 2, January 2023

A Recent Survey Paper on Text-To-Speech Systems

Shruti Mankar[1], Nikita Khairnar[2], Mrunali Pandav[3], Hitesh Kotecha[4],Manjiri Ranjanikar[5]

Pimpri Chinchwad College of Engineering, Nigdi, Pune, Maharashtra, India[1,2,3,4,5]

Abstract: *All of us are aware of how important knowledge is and it is also true that mostly the data or knowledge is in the form of books or online articles, or various pdfs i.e. in text format. But not all of us are privileged to read. Some are illiterate, some are blind, and some have reading difficulties. Hence, a form of adaptive technology or procedure that reads digital text aloud, which is called text-to-speech (TTS) was developed. It is occasionally referred to as "read-aloud" technology. Words on a computer or other digital device can be converted into audio using TTS. TTS is particularly beneficial for all those people who have reading difficulties or are illiterate. Also, a significant amount of research has been done and is currently being done on text-to-speech technology. Various technologies, methodologies, and algorithms are used in the various proposed approaches and solutions for TTS. This research presents a systematic review of all those methods which have been proposed and implemented by different active researchers in this field.*

Keywords: Text-to-speech conversion, text-to-speech synthesis, machine learning, neural networks, optical character recognition, etc

I. INTRODUCTION

Text-to-voice conversion basically deals with the technology of converting any type of character text to voice signals. In other words, it is a speech manufacturer that mouths the word in a realistic way. This technology has become very popular among physically impaired people. Apart from this it also helps to increase the literacy rate. There are various applications for TTS technology such as tutorials for automobile driving, reading books, and speeding up the task. It has also been used for making voice assistants such as Google Assistant, Alexa, etc.

Converting text to speech is typically conceived of as a two-step process. The text is initially transformed into an underlying abstract linguistic representation made up of phonemes, stress symbols, and signs of grammatical structure markers. Then, using a set of rules that control a voice box model the sequence of phonemes is transformed into sound.

Basically, the focus is on surveys to prepare a functional, efficient, and real-time beneficial system that helps people read a newspaper to the people. This not only increases literacy among the uneducated but also speeds up the task as it will reduce the time to read paper manually.

The next sections of the paper are a literature survey that covers various proposed methods, then a conclusion and references.

II. APPLICATION OF TTS

1. **Telecommunication**: Text-to-speech can produce words from a customer's data that are read back to them in a nice, professional voice and can deliver personalized messaging that the caller can interact with.
2. **Banking and Finance**: Text-to-speech can be used to increase security and improve the customer experience by making it more accessible, dynamic, and personalized. It also gives you the freedom to check your money and the stock market on the go using nothing more than voice commands.
3. **Travel and Tourism**: Self-guided audio tours powered by artificial voices can be created using text-to-speech software.

DOI: 10.48175/IJARSCT-7954

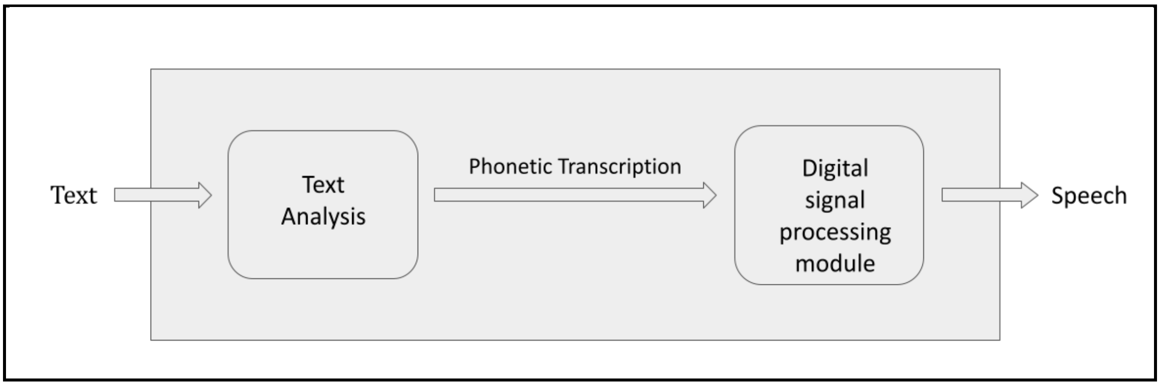

Figure 1: Generalized system architecture of text-to-speech conversion system

III. LITERATURE REVIEW

This literature review is founded on different methods and approaches that have been proposed to develop an efficient text-to-speech system. We have studied those proposed methods, their advantages and disadvantages, algorithms used to develop those systems, and some more details about the proposed approaches. The literature review done by us is as below.

Mrunmayee Patil et al., 2014 [1] with the goal to recommend a method or system for converting audio to text and images to text that created a convenient user interface for blind persons wrote a paper titled *“A review on conversion of image to text as well as speech using edge detection and image segmentation”*. The suggested framework implemented an image-to-text and then used edge detection and the Canny method. Used edge detection algorithm, generates output that already emphasizes the position of edges with a high value, resulting in extremely thin and clean edges. But it takes a long time. The outcome of this research may be summed up as follows: The proposed study has given greater weight to object recognition in images. This finally leads to recognizing significant items in a picture.

Itunuoluwa Isewon et al., 2014 [2] proposed the technique suggested in the paper *“Design and implementation of text to speech conversion for visually impaired people”* which reads aloud any text that has been entered and can also be stored as an mp3 file after being converted to voice. The suggested approach uses both Digital signal processing (DSP) technology and the NLP i.e., natural language processing techniques. The program's name is "TextToSpeech Robot" (TTSR), and it was created in Java. Text can be entered into the text box provided for TTSR or copied and pasted from an external document onto the local computer to convert it to voice. Users of TTSR can transfer the audio format of previously converted text to any of their audio devices by saving the audio file in any spot on their local PC.

Venkateswarlu et al., 2016 [3] proposed a system in the paper entitled *“Text to speech conversion”*. The proposed system of TTS and SST can be implemented for different languages depending upon the user’s requirement and has the capability to recognize these languages and change them to the desired text or speech format. It has provided two functionalities i.e send an email by STT**,** Convert TTS at the receiver’s end. Systems can be deployed for the purpose of effective communication among illiterate and visually challenged people. HMM is a statistical model therefore most suitable for both STT and TTS conversions.

Shuang Ma et al., 2019 [4] proposed a brand-new generative model for the development of a skip-modality in a paper titled *"Unpaired image-to-speech synthesis with multimodal information bottleneck"*. This is in contrast to deep generative models, where usually paired data with exact correlation across the modalities is needed for training these models. In order to execute the suggested method, picture text pairings and text audio pairs were employed as common modalities which utilized a 3-layer convolutional neural network, i.e. CNN and maximum pooling. Hence the suggested technique successfully shows image-to-speech synthesis in the absence of paired data.

Tae-Ho Kim et al.,(2019) [5] suggested a paper entitled *"Emotional Voice Conversion Using Multitask Learning With Text-To-Speech"* where a voice converter using multitask learning is presented. Text-to-speech conversion is a technique to transform textual information into speech waveform. The work is expanded upon for emotional voice conversion in the article. The style encoder in this method captures emotional data. Therefore, the paper's suggested model can handle many-to-many emotional voice conversions. The voice conversion and text-to-speech conversion tasks can both be completed by the model that is currently being used in the study. The paper's key contribution was its use of sequence-to-sequence models to achieve many-to-many emotive voice conversions. Multitask learning using TTS can also boost VC performance. The conclusion can be drawn because multitasking encourages subjective assessment. Korean parallel databases were used to test the data. Additionally, the encoder is taught to solely extract emotional data rather than grammatical data.

Cong Zhou et al., (2019) [6] suggested a paper entitled *"Voice Conversion with Conditional SampleRNN"* where a novel approach is presented for voice conversion based on conditioning the SampleRNN. The voice material will be preserved using this method. The SampleRNN model for multiple speakers was initially trained. The suggested approach can convert numerous voices at once. The study mentions the deep autoregressive model approach after examining the issues with conventional voice converting. The strategy is highly efficient and easy to use. It necessitates regular training on parallel datasets for a conventional system. The characteristics are retrieved from the signals during this training. There might be a number of errors during this entire period. Deep neural networks have lately been used for voice conversion as well. But enhancing speech quality is the major objective. The study suggests the usage of SampleRNN, a generative model for producing high-quality audio.

Kuan Chen et al.,2018 [7] suggested a paper entitled *"High-quality Voice Conversion Using Spectrogram-Based WaveNetVocoder"* where a WaveNet model based on spectrogram is proposed. In this, the source speaker is converted to the target speaker using LSTM RNN-based frame-to-frame feature mapping. In comparison to conventional methods, the WaveNet waveform generator based on spectrograms provides audio of higher quality. Parallel data was employed in traditional voice converter methods. The most crucial factor in voice conversion is sound quality. Acoustic characteristics used in conventional techniques involve vocoder parameters, whose conversion might result in quality distortion. The waveform-generating method that can create high-quality voice waveforms is WaveNet. The author of this study suggests a mel-frequency spectrogram-based high-quality voice converter architecture. Using a vocoder, these properties are vocoded into waveforms. According to the article, voice conversion using Mel-frequency spectrograms can result in high-quality voice.

Shivangi Nagdewani et al., 2020 [8] presented a paper titled *"A Review On Methods For Speech-To-Text And Text-To-Speech Conversion"*. They put forth a creative, cost-effective method that allows users to hear rather than read the contents of text graphics. There are two modules: one deal with image processing, which mostly makes use of OCR technology, and the other deals with voice processing, which primarily makes use of TTS technology. We can simplify the editing process for books and websites by using this method. Through a vocal interface, this device enables persons who are visually challenged to connect with computers successfully.

Nanxin Chen et. al. (2021) [9] suggested a paper titled *"WaveGrad 2: Text-to-Speech Refinement Through Iteration"* in which he wrote a generative model that is non-autoregressive for text-to-speech synthesis. WaveGrad 2 is trained using a phoneme sequence to estimate the gradient of the waveform's log conditional density. The model constructs an audio waveform using an input phoneme sequence and an iterative refining procedure. In contrast, the original WaveGradvocoder relied on Mel-spectrogram features created by a different model. Starting with Gaussian noise, the iterative refinement procedure finally recovers the audio sequence across a number of refinement stages (for example, 50 steps). WaveGrad 2 provides an easy solution to balance inference efficiency and sample quality by adjusting the number of refinement stages. Experiments reveal that the model can generate high-fidelity audio that is comparable to the capabilities of a cutting-edge neural TTS system.

Isaac Elias et. al. (2021) [10] suggested a paper titled *"Parallel Tacotron 2: A Non-Autoregressive Neural TTS Model with Differentiable Duration Modeling"* in which he offers Parallel Tacotron 2, a non-autoregressive neural text-to-speech model with a totally differentiable duration model that does not need supervised duration signals A novel attention mechanism and an iterative reconstruction loss based on Soft Dynamic Time Warping are used to build the duration model. It can learn token-frame alignments and token durations on its own. Parallel Tacotron 2 exceeds baselines in terms of perceived naturalness across a number of multi-speaker evaluations, according to experimental data. It also demonstrates its capacity to manage duration.

KainanPenget. al. (2020) [11] suggested a paper titled *"Non-Autoregressive Neural Text-to-Speech"* in which they suggest ParaNet, a non-autoregressive text-to-spectrogram seq2seq model. When compared to the lightweight Deep Voice 3, it speeds up speech synthesis by 46.7 times while preserving passably respectable voice quality. On the challenging test sentences, ParaNet consistently aligns text and speech by gradually improving attention through

DOI: 10.48175/IJARSCT-7954

iteration. We also build a parallel text-to-speech system and test a number of parallel neural vocoders that can convert text into voice in only one feed-forward pass. We also look into a novel VAE-based technique that starts from scratch and trains the parallel inverse autoregressive flow vocoder rather than distilling from a separate WaveNet.

Jeff Donahue et. al. (2021) [13] suggested a paper titled *"End-to-End Adversarial Text-to-Speech"* in which he used Multiple processing steps are frequently used in contemporary text-to-speech synthesis pipelines, each of which is created or learned independently of the others. This study addresses the difficult problem of constructing models that operate directly on character or phoneme input sequences and create raw speech audio outputs by learning to synthesize speech from normalized text or phonemes in an end-to-end way. Our suggested generator is feed-forward and uses a differentiable alignment technique based on token length prediction, making it suitable for both training and inference. Using a mix of adversarial feedback and prediction losses, it learns to create high-fidelity audio while keeping the produced audio confined to closely approximate the ground truth in terms of total time and Mel-spectrogram.

Sr No.	Dataset Name	Link to database	Paper
1.	The LJ Speech dataset	https://keithito.com/LJ-Speech-Dataset/	fastpitch: parallel text-to-speech with pitch prediction [15]
2.	OGI diphone databases	http://festvox.org/bsv/c2265.html	spectral voice conversion for text-to-speech synthesis [16]

Table 1: Dataset used by authors

IV. PROPOSED SYSTEM ARCHITECTURE

The proposed system will be able to convert text to speech. The user must enter the text, which will then be turned into speech. It allows you to change the pitch and rate (speed) of the speech output. Aside from that, the user will be able to choose the accent in which all the voice output would be delivered.

The below Fig 2 shows the flow diagram representation of the proposed system.

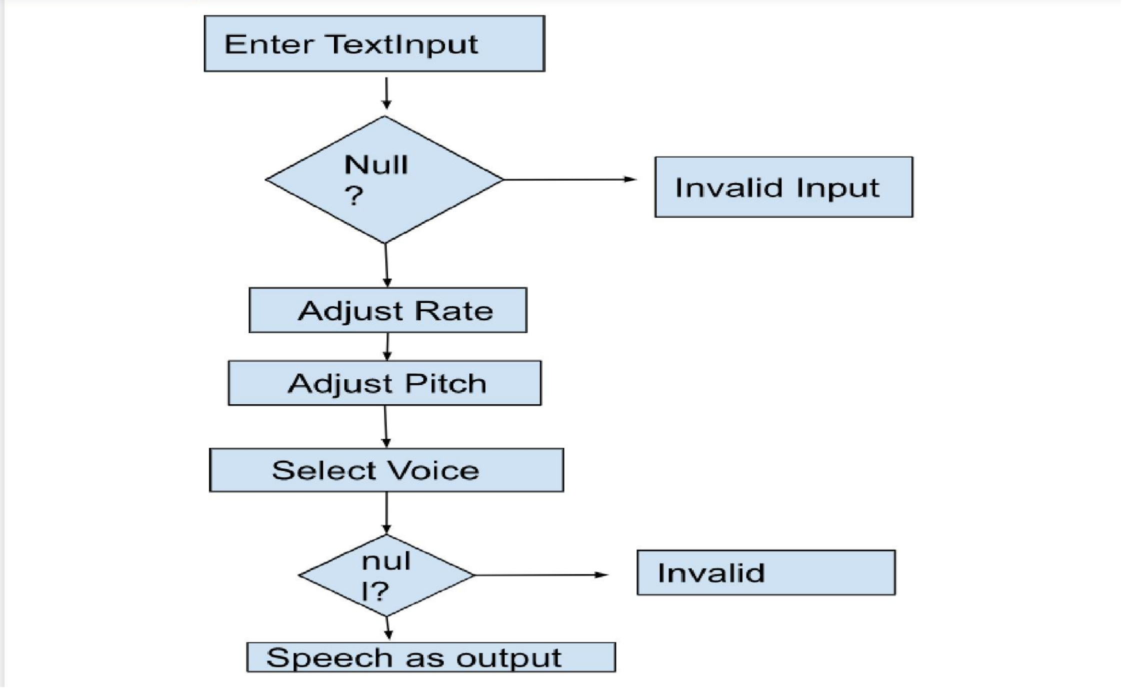

Figure 2: Flowchart for Text-to-speech Synthesis App

DOI: 10.48175/IJARSCT-7954

V. FUTURE SCOPE

Evaluation based on these research papers shows that some extensions are present for web browsers to convert the web pages' text to voice. Also, there are a number of applications and methods which help with TTS and STT. But if we talk about a system that will convert web pages and any form of text to a voice signal. There is a requirement for such a system. These advantages of the existing system can be merged to develop a single system that will provide such functionality. That system can be deployed for the plan of effective communication by typical, unlettered, and partially-sighted people.

VI. CONCLUSION

Text-to-speech, often known as TTS, is a type of assistive technology that allows a computer or tablet to read aloud the words that are displayed on the screen to the person who is using it. Students who have issues reading, particularly those who have trouble decoding, are big fans of this technology. The pupil is able to concentrate on the meaning of the words rather than expending all of their mental energy attempting to sound out the words when the words are presented to them in an auditory format. Despite the fact that it can aid students in overcoming their reading issues and accessing the content presented in the classroom, this technology does not help pupils improve their reading abilities.

This paper shows the literature review that is founded on different methods and approaches that have been proposed to develop an efficient text-to-speech system. We have studied those proposed methods, their advantages and disadvantages, algorithms used to develop those systems, and some more details about the proposed approaches.

REFERENCES

[1]. Patil Mrunmayee and Ramesh Kagalkar, "A review on conversion of image to text as well as speech using edge detection and image segmentation", International Journal of Advanced Research in Computer Science Management Studies 2, 2014

[2]. Isewon, Itunuoluwa, Jelili, Oyelade, and OlunfunkeOladipupo, “Design and implementation of text to speech conversion for visually impaired people”, International Journal of Applied Information Systems 7, no. 2, pp. 25-30, 2014

[3]. Venkateswarlu, S., D.B.K. Kamesh, J.K.R. Sastry, and Radhika Rani, “Text to speech conversion”, Indian Journal Of Science and Technology 9, no. 38, pp. 1-3, 2014

[4]. Ma, Shuang, Daniel McDuff, and Yale Song, “Unpaired image-to-speech synthesis with multimodal information bottleneck” In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7598-7607, 2019

[5]. Tae-Ho Kim, Sungjae Cho, Shinkook Choi, Sejik Park? andSoo-Young Lee ."Emotional Voice Conversion Using Multitask Learning with Text-To-Speech" ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)(2019).

[6]. Cong Zhou, Michael Horgan, Vivek Kumar, Cristina Vasco, Dan Darcy. "Voice Conversion with Conditional SampleRNN" Interspeech 2018, Hyderabad, India.

[7]. Kuan Chen, Bo Chen, Jiahao Lai, Kai Yu. " High-quality Voice Conversion Using Spectrogram-Based WaveNetVocoder" Interspeech, 2018.

[8]. Nagdewani, Shivangi, and Ashika Jain. "A Review On Methods For Speech-To-Text And Text-To-Speech Conversion " International Research Journal of Engineering and Technology, (2020).

[9]. Donahue, J., Dieleman, S., Bińkowski, M., Elsen, E., &Simonyan, K. (2020). End-to-End Adversarial Text-to-Speech.arXiv. https://doi.org/10.48550/arXiv.2006.03575

[10]. T. -H. Kim, S. Cho, S. Choi, S. Park and S. -Y. Lee, "Emotional Voice Conversion Using Multitask Learning with Text-To-Speech," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 7774-7778, doi:10.1109/ICASSP40776.2020.9053255.

[11]. Kuan Chen, Bo Chen, Jiahao Lai, Kai Yu, "High-quality Voice Conversion Using Spectrogram-Based WaveNetVocoder," Interspeech 2018- isca-speech.org.

[12]. MingyangZhang,Xin Wang2, Fuming Fang2, Haizhou Li1, Junichi Yamagishi2, "Joint training framework for text-to-speech and voice conversion using multi-source Tacotron and WaveNet," April 2019, doi: https://arxiv.org/abs/1903.12389.

[13]. Tae-Ho Kim, Sungjae Cho, Shinkook Choi, Sejik Park and Soo-Young Lee, "Emotional Voice Conversion Using Multi task Learning with Text to Speech " , November 2019, arXiv:1911.06149v2.

[14]. Cong Zhou, Michael Horgan, Vivek Kumar, Cristina Vasco, Dan Darcy, "Voice Conversion with Conditional Sample RNN", Interspeech 2018, Hyderabad, India. https://doi.org/10.48550/arXiv.1808.08311.

[15]. Łańcucki, Adrian, "Fastpitch: Parallel text-to-speech with pitch prediction" In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6588-6592, IEEE, 2021

[16]. A. Kain and M.W. Macon, "Spectral voice conversion for text-to-speech synthesis", Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat.No. 98CH36181), 1998, pp. 285-288 vol. I, doi: 10.1109/ICASSP.1998.674423