RESEARCH ARTICLE | SEPTEMBER 05 2023

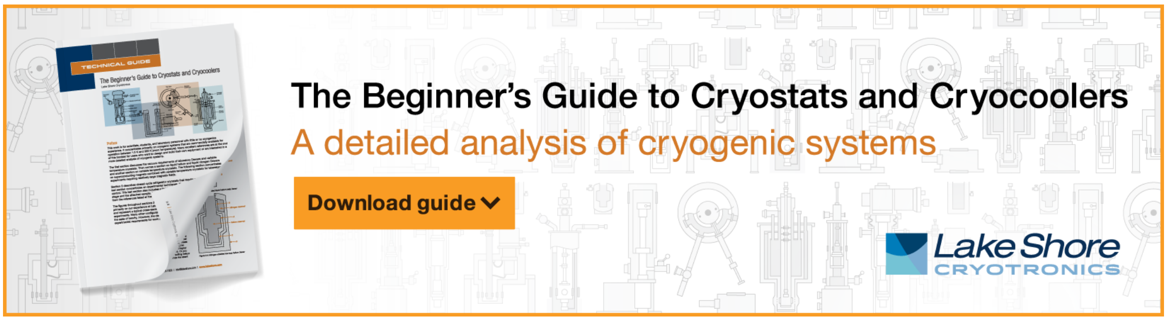
Custom tone generator

T. V Suneetha; Mitlamalla Nikhil; Siripuram Rohith; Arvapally Rahul; Kajana Uday

Check for updates

AIP Conf. Proc. 2754, 140001 (2023)
https://doi.org/10.1063/5.0166908

View Online
Export Citation
CrossMark



Custom Tone Generator

T.V Suneetha[1,a)], Mitlamalla Nikhil[1,b)], Siripuram Rohith[1,c)], Arvapally Rahul[1,d)], Kajana Uday[1,e)]

[1]*Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, Telangana, 500090 India.*

a) Corresponding author: takkellapati9@gmail.com
b) mitlamallanikhil123@gmail.com
c) rohith.siripuram20@gmail.com
d) rahularvapally@gmail.com
e) kajanauday@gmail.com

Abstract. People don't have enough time to read their books, and many people struggle to do so. Some folks had grown tired of hearing the default voices. We have devised a solution called Custom Tone Generator to address these issues. Custom Tone Generator is an application that turns documents into personal audio files with the help of optical character recognition so that you can better evaluate the data. We will then convert the document to an audio file using speech models that have been trained. To continue, we collect a sample of the user's tones and then train the model. We also have an option to save the trained model. Once the trained model has been created, it is used to turn the text document into an audio file with the required tone. Then the user can listen to the audio. This application allows the user to listen to the content of any book in their preferred tone, thereby eliminating the need to spend time and effort reading books. Users can listen to the books and do whatever work they want while listening.

INTRODUCTION

In this digitalized era, mobile phones have been widely utilized for communication for the past few years. It is simple to converse with one another via phone conversations and text messages. The most appropriate mode of transmitting and receiving precise information is verbal communication. Engage with local and distant services to better assist the people. Text-to-speech was originally created to assist the visually impaired by providing a computer-generated spoken voice that people could hear by reading the text. The text in documents or images of typed-to-speech conversion will be examined in this project. Character recognition is a method for converting printed letters and digits into digital text. OCR is used in our project. OCR reads characters from a scanned document or typed image, processes them, extracts features, and recognizes patterns. We must first pre-process the typed picture before using OCR. To make the image better to handle, pre-processing includes compression, noise removal, and thresholding. Character recognition relies heavily on segmentation and feature extraction processes. The image is segmented into several parts throughout this process. This aids in the segmentation of each character in the image of the typed text. OCR translates viewable photos into text files that can then be converted to speech. Readable photos are those with less intricacy in the background, allowing the letters to be extracted grammatically. The user must provide a typed or document image as input, which will be processed and the content in the image saved in a text file. The user has the option of altering the resulting text file. The database stores the created text files. A speech synthesis system is then used to transform the text file into speech. The machine learning model is implemented in an Android application. Our project is useful in assisting blind and visually impaired people.

REVIEW OF LITERATURE

Knowledge extraction by just listening to sounds is a distinctive property. Speech signals are a more effective means of communication than text because blind and visually impaired people can also respond to sounds. This paper aims to develop a cost-effective, and user-friendly optical character recognition (OCR) based speech synthesis system. The OCR-based speech synthesis system has been developed using laboratory virtual instruments and engineering workbenches. [1]

Natural language processing is a widely used technique by which systems can understand the instructions for manipulating text or speech. In the present paper, a Text-to-speech synthesizer is developed that converts text into spoken words, by analyzing and processing it using Natural Language Processing (NLP) and then using Digital Signal Processing (DSP) technology to convert this processed text into a synthesized speech representation of the

text. Here we developed a useful text-to-speech synthesizer in the form of a simple application that converts inputted text into synthesized speech and reads it out to the user, which can then be saved as an MP3 file. [2] Text Reader for Blind

International Conference on Materials, Processing & Characterization
AIP Conf. Proc. 2754, 140001-1–140001-6; https://doi.org/10.1063/5.0166908
Published by AIP Publishing. 978-0-7354-4634-2/$30.00

People using a camera module ensuring portability is the prototype made using the Raspberry Pi 3b and Python to read the text from the handheld objects of the blind person. This paper proposes a better approach for text localization and extraction for the detection of text areas in images. The text size is an important factor whose dimension should be properly elected to make the method more general and insensitive to various font shapes and sizes. [3]

Visually impaired people are dependent solely on Braille books and audio recordings provided by NGOs. Owing to many constraints in the above two approaches, blind people can't have the book of their choice. The presented work will provide them an opportunity to have an audiobook of their choice in English or Marathi, or any printed book having English, Marathi, or Braille script. Printed text from textbooks having English, Marathi, or Braille script will be taken as input in the form of an image, which will be converted into plain editable text with the help of optical character recognition (OCR). This plain text will be then fed to the Text to Speech (TTS) converter, which will generate the audio output file in English or Marathi corresponding to the input text image script. A printed book to audiobook converter has been successfully implemented, and satisfactory results were obtained. [4]

RESEARCH METHODOLOGY

Construction and Implementation Details

The proposed approach consists of the following modules:

Authentication

Authentication for users is provided by using Firebase Backend Services and easy-to-use SDKs. It accepts passwords, phone numbers, and prominent federated identity providers like Google, Facebook, and Twitter, among other methods.

Database

Firebase is the database that we have used for this project. In Firebase, we primarily use Cloud Fire Store and Cloud Storage. Cloud Fire Store is a versatile, scalable database from Firebase and Google Cloud for mobile, Web, and server applications. It's a cloud database that uses NoSQL technology. Firebase Cloud Storage is a robust, easy-to-use, and cost-effective object storage service.

Text Recognition

Text recognition is a technique for extracting text from photos or documents. As an output, it creates a dot text file. The file (.txt) is saved in cloud storage.

Text-to-speech

The text-to-speech converter generates voice from a dot txt file obtained through Cloud Storage.

Custom Tone Generator

Custom tones are trained in this module. Finally, the text document is translated into the pre-trained preferred voice.

Data Set

We will take the document and an audio file as input. For ML Kit to precisely perceive text, input pictures should contain text addressed by adequate pixel information. Preferably, each character ought to be essentially 16x16 pixels. There is, for the most part, no precision advantage for characters to be bigger than 24x24 pixels. Thus, for instance, a 640x480 picture may function admirably to examine a business card that possesses the full width of the picture. To check a record printed on letter-sized paper, a 720x1280 pixel picture may be required. Poor image focus can affect text recognition accuracy.

This app accepts documents of less than or equal to 100 pages because, as the machine learning model is running on a mobile device, it occupies a lot of RAM. If the document is more than 100 pages, the app may crash; to avoid this, we set a standard input size of 100 pages. We have used LibriSpeech/train-clean-100 dataset. The Custom Tone Generator is a deep learning framework in three stages. In the first stage, one creates a digital representation of a voice from a few seconds of audio. In the second and third stages, this representation is used as a reference to generate speech given arbitrary text.

System Architecture

Figure 1 depicts the system architecture of the suggested approach. It depicts the system's whole physical deployment. The document that is sent can be in any format. After scanning and converting the image to a text file, it is pre-processed and saved in the user database. The text-to-speech converter, which is OCR, converts the text file into an audio file when it is added, where the user has the ability to change the text file's audio tone.

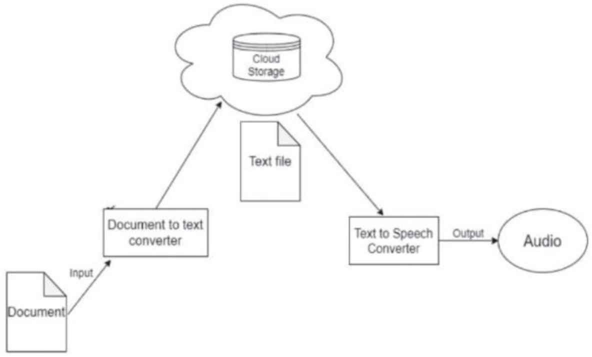


FIGURE 1. System Architecture

The proposed approach is implemented in the four steps below.

Step 1- Authentication:

In this activity, the user can select a provider by clicking on the LOGIN/REGISTER button, then clicking on the provider that he or she wants to use and entering the credentials to log in or register.

Step 2- Document or image-to-text conversion:

In this activity, the user can select an image to convert to text by clicking on the camera button, and then select the document by clicking on the PDF button. After you've made your selection, click the Convert button to convert to text.

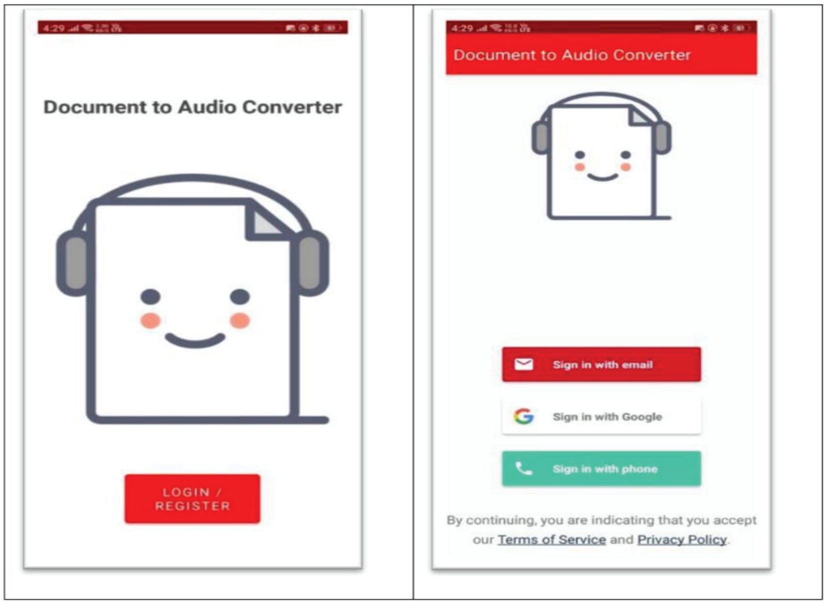
Step 3- Files that have been converted by the user:

All converted user files from the Fire Store are presented in this activity, and users can click on a file to listen to it. There is also a floating button that allows the user to convert a typed or document image to text by simply clicking on it.

Step 4- Conversion of text to speech:

The Text-to-Speech engine is launched in this activity, and after it is invoked, the user can vary the pitch and speed of the voice, as well as switch to a male or female voice, and the text is read by the Text-to-Speech engine by pressing the play button and can generate output in the custom voice.

RESULTS



By continuing, you are indicating that you accept our Terms of Service and Privacy Policy.

Figure 2. Authentication

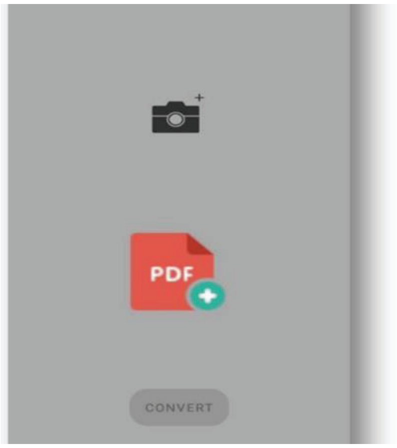
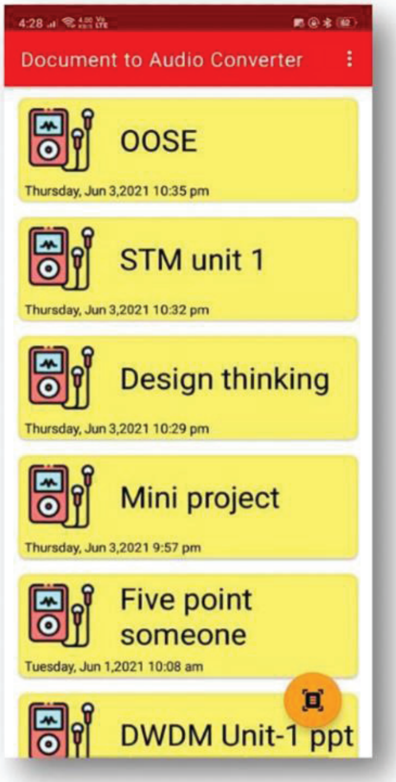


Figure 3. Document or Image to Text Conversion



4:28

Document to Audio Converter

OOSE

Thursday, Jun 3,2021 10:35 pm

STM unit 1

Thursday, Jun 3,2021 10:32 pm

Design thinking

Thursday, Jun 3,2021 10:29 pm

Mini project

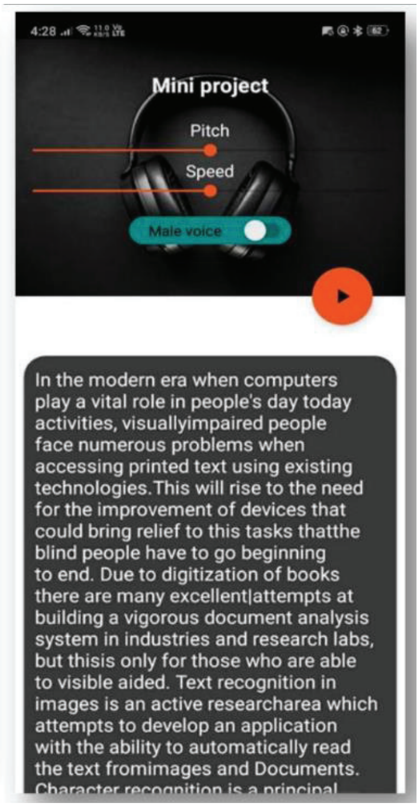
Thursday, Jun 3,2021 9:57 pm

Five point someone

Tuesday, Jun 1,2021 10:08 am

DWDM Unit-1 ppt

Figure 4. User Converted Files

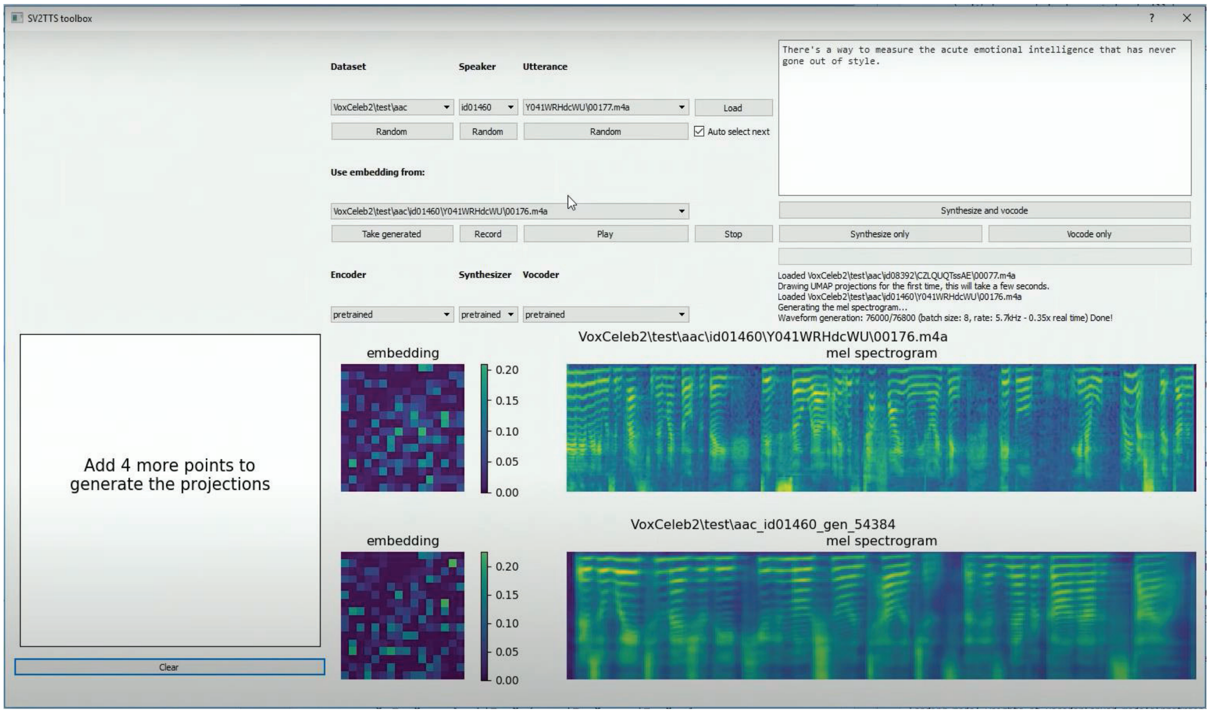


4:28

Mini project

In the modern era when computers play a vital role in people's day today activities, visuallyimpaired people face numerous problems when accessing printed text using existing technologies.This will rise to the need for the improvement of devices that could bring relief to this tasks thatthe blind people have to go beginning to end. Due to digitization of books there are many excellent|attempts at building a vigorous document analysis system in industries and research labs, but thisis only for those who are able to visible aided. Text recognition in images is an active researcharea which attempts to develop an application with the ability to automatically read the text fromimages and Documents. Character recognition is a principal

Figure 5. Conversion of Text to Speech



SV2TTS toolbox

Dataset: VoxCeleb2\test\aac

Speaker: id01460

Utterance: Y041WRHdcWU\00177.m4a

☑ Auto select next

Use embedding from: VoxCeleb2\test\aac\id01460\Y041WRHdcWU\00176.m4a

Encoder: pretrained

Synthesizer: pretrained

Vocoder: pretrained

```
There's a way to measure the acute emotional intelligence that has never
gone out of style.
```

```
Loaded VoxCeleb2\test\aac\id08392\CZLQUQTssAE\00077.m4a
Drawing UMAP projections for the first time, this will take a few seconds.
Loaded VoxCeleb2\test\aac\id01460\Y041WRHdcWU\00176.m4a
Generating the mel spectrogram...
Waveform generation: 76000/76800 (batch size: 8, rate: 5.7kHz - 0.35x real time) Done!
```

Add 4 more points to generate the projections

Figure 6. Backend System for Custom Tone Generation

LIMITATIONS

ξ This application supports only the English language.
ξ Model performance may be reduced if the document contains the undecipherable text.
ξ The accuracy of the model is affected if the background noise changes continuously.

CONCLUSION

Character recognition is a difficult problem that will take a long time to solve. The requirements for recognizing characters revolve around datasets. This machine learning model is designed to take an image or a document as input, detect the text, and generate a voice from it. This application can be used in a variety of fields, including education. This application converts scanned documents to text in a matter of seconds. Characters are detected with a high degree of precision. This software assists a child in listening to stories. It also makes it easier for blind people to read books.

FUTURE SCOPE

In the future, this project can be extended to include accent and language customizations that the user wants to listen to and can be more focused on various custom tones.

REFERENCES

1. J.J.Mullani, M.Sankar, P.S.Khade, S.H.Sonalkar and N.L.Patil, "OCR BASED SPEECH SYNTHESIS SYSTEM USING LAB VIEW: Text to Speech Conversion System using OCR", in *ICCMC*,(2018), pp.7- 14.
2. P.M.ee, S.Santra, S.Bhowmick, A.Paul, P.Chatterjee and A.Deyasi,"Development of GUI for Text-to-Speech Recognition using Natural Language Processing," in *IEMENTech* , (2018).
3. T.Shah and S. Parshionikar,"Efficient Portable Camera Based Text to Speech Converter for Blind Person," in *ICISS*, (2019).
4. A.Domale, B.Padalkar, R.Parekh and M.A.Joshi, "Printed Book to Audio Book Converter for Visually Impaired," in *Texas Instruments India Educators Conference* (2013), pp.114-120.
5. Chunyong Ma, Anni Wang, Ge Chen and Chi Xu, *The Visual Computer*, (*Springer Link* ,2018). 6. Nasser H. Dardas and Nicolas D. Georganas, “Real-Time Hand Gesture Detection and Recognition Using Bag-of-Features and Support Vector Machine Techniques,” (IEEE, 2011).
7. Tao Liu, Wengang Zhou and Houqiang Li, “Sign language recognition with long short-term memory” ,in *IEEE International Conference on Image Processing*, (2016), pp.2871-2875.
8. Arpita Haldera and Akshit Tayade, *International Journal of Research Publication and Reviews* (2021).