Text-to-Speech Recognition using Google API

Orlunwo Placida Orochi
Computer Science Department
Ignatius Ajuru University of Education

Ledisi Giok Kabari
Computer Science Department
Ignatius Ajuru University of Education

ABSTRACT

Speech is the most natural mode of human communication. To enable machines to understand human speech, computers can act as an intermediary for human experts, allowing them to respond accurately and reliably to human voices. This can be accomplished by a text-to-speech recognition device, which allows a data processor to accurately interpret the language in which a message was written and translate it to an audio file that can be heard through a sound medium such as a speaker. This paper uses Python programming language to introduce a text-to-speech model to see whether the messages written are read. Using Google API, text-to-speech conversion was successful.

Keywords

API, Artificial Intelligence, Speech-to-speech

1. INTRODUCTION

Speech Recognition is part of natural language, an artificial intelligence subfield. Thus, speech recognition is a computer program that identifies and converts words and phrases to human readable text in a spoken language. It is used in various applications including voice support, home automation, voice-based chatbots, robot communicating voices, artificial intelligence, etc.

A text to speech synthesizer (TTS) is a device based on a computer which automatically read the text out aloud irrespective of whether the text is inserted by a computer input stream or by an OCR-engine. Both hardware and software can implement a speech synthesizer. Speaking is also based on the concatenation of natural language, i.e., units taken from natural language forming a word or phrase.

Real-time streaming of speech to text: The API can process audio in real-time from the system microphone or use an audio file as a source and even translate it into text.

Various domain-based models: Depending on the project specifications, you can choose between various qualified models. For instance, the enhanced call model can be used to convert audio from a telephone.

Adaptation: You can modify an API by turning it into additional classes to understand rare terms, currency, numbers etc.

2. RELATED WORKS

Numerous language translators are available to enter the content manually and the app is working to ensure that the content is translated into the correct language. For this method, some translation platforms also fee. For instance, the most common Google API translation fees calculated by use [1].

The proposal for the language recognition scheme based on the techniques MFCC and VQ was made by [2]. Higher accuracy for more languages has been achieved by using MFCC. This system was primarily used to classify the language without speaking or speaking acknowledgment.

[3] suggested a speech check system using MFCC and a function matching and speaker modeling system using a Vector Machine (SVM). The precision of the method is compared by different orders of the use of the MFCC coefficients. The findings indicate a higher precision of 20 to 25 MFCC coefficients. Speech to text was not integrated and authentication was given and not identification was provided.

The proposed Voice Reconnaissance System for security purposes proposed by [4,5]. Which uses MFCC and VQ for modeling functions. The system's downside is that it relies on text. The system only recognizes speakers and does not involve translation of speech to text [2]. A technique for the extraction of feats for Arabic speech recognition systems was suggested by [6]. To do this, they used MFCC.

[7] supervised and uncontrolled voice recognition approaches with LSTM and restrained Boltzmann machines on recurrent neural networks respectively. A literary analysis on automatic speech recognition was conducted by [8]. ASR history, classification and speech recognition methodologies have been addressed. [9] has suggested the use without the inclusion of RNNs of connectionist temporal classification with CNN. This kind of model can be calculated with considerable precision. An RNNs, LSTMs and their output contributions for voice recognition systems have been provided by [10].

In a speech age classification system that evaluates the performance of the GMM grade level 1 and 2 was proposed by [11]. The two level GMM was designed with different mixes and two level GMM was shown to be more accurate than a level GMM, but in two levels GMM based method, the calculation delay was greater. Speaking recognition system text output can be sent either one-in-one or many-in-one to other computers.

The text-independent speaker verification method was implemented by [12,13] where the power spectrums were used along with GMM to increase system accuracy. [14] proposed to use GMM-based clustering algorithms for the identification of the speaker as function matches. In this case, the inconvenience of k-mean clustering is overcome by using the initial algorithm that uses T-test matrices to identify the distance. This increased precision.

3. METHODOLOGY

Using similar technological advancement found in signal processing of which [15] were able to detect similarities in sound, our study aim is to confirm that text signals are properly translated using Text-To-Speech; a method that first analyzes, then processes and understands the inputted texts, then converts to digital audio and speaks in the language for which the text was made. The block scheme of TTS is shown in Fig. 1. As such the proposed model Fig. 1, is a reverse application of signal processing as against what was done in the study [15].

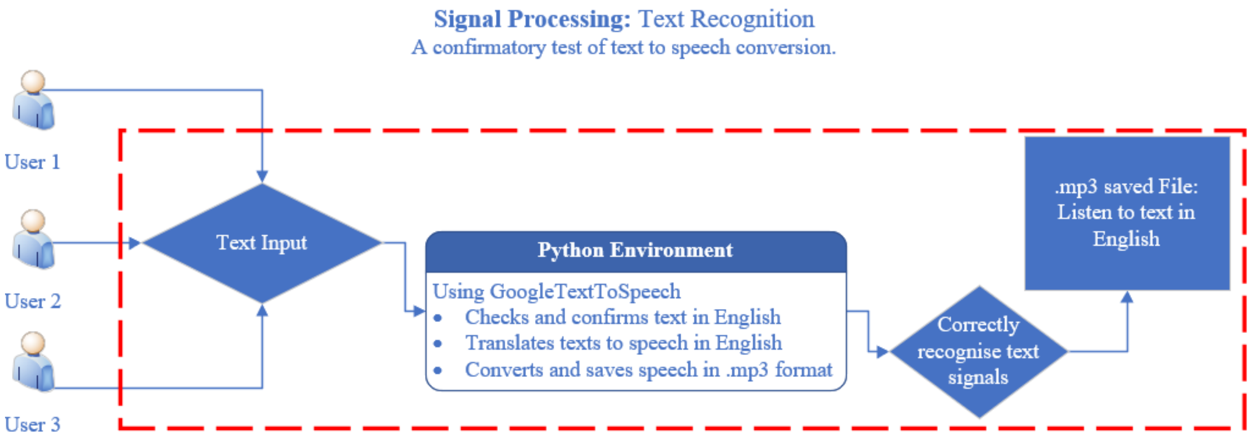

Signal Processing: Text Recognition
A confirmatory test of text to speech conversion.
User 1
User 2
User 3
Text Input
Python Environment
Using GoogleTextToSpeech
Checks and confirms text in English
Translates texts to speech in English
Converts and saves speech in .mp3 format
Correctly recognise text signals
.mp3 saved File: Listen to text in English

Fig 1: Architecture of Text-to-Speech

4. RESULT AND DISCUSSION

Several APIs are available for Python conversion of text to voice. The Google Text to Speech API, usually referred to as the gTTS API, is among these APIs. gTTS is a tool that transforms entered text into an mp3 format that can be saved as an audio tool.

The gTTS API supports several languages such as English, Hindi, Tamil, German, French and many others. The speech can be transmitted quickly or slowly at one of the two audio speeds available. However, for this study only texts in English are read and translated into speech in English .mp3 audio format. of the audio produced cannot be changed from the latest update.

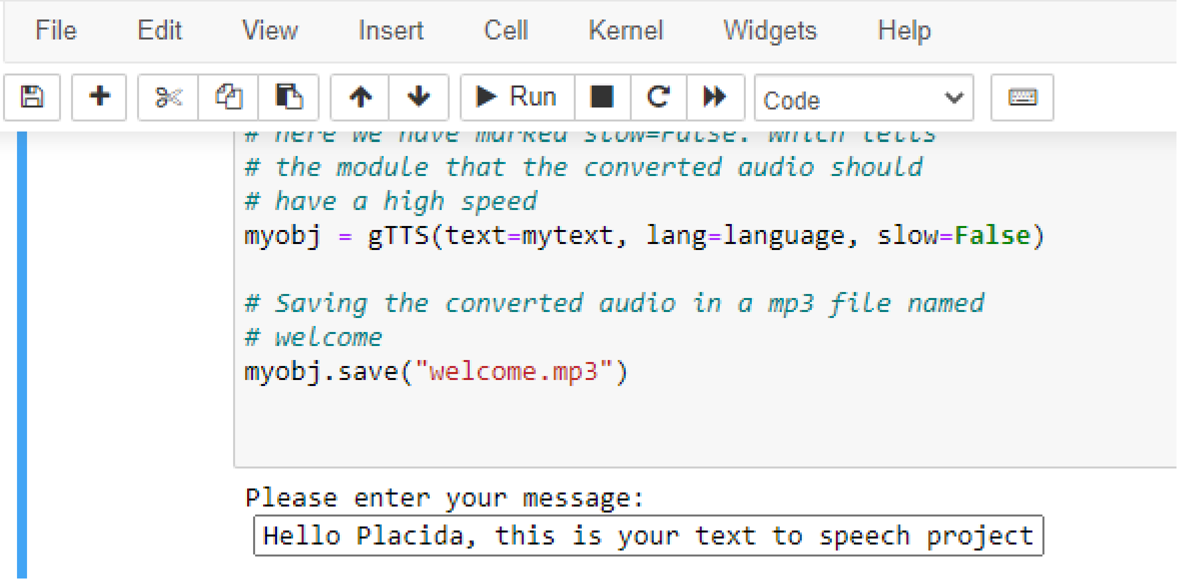
Observed outcome of project in Fig. 2:

i. When python code runs successfully, user is required to input a message (text).

ii. User hits the enter key

iii. Google API recognises the language (English) for which the message is written

iv. Message is translated to English and converted into audio file which is saved as .mp3 format.



```
# here we have marked slow=False. Which tells
# the module that the converted audio should
# have a high speed
myobj = gTTS(text=mytext, lang=language, slow=False)

# Saving the converted audio in a mp3 file named
# welcome
myobj.save("welcome.mp3")
```

```
Please enter your message:
Hello Placida, this is your text to speech project
```

Figure 2: User Message Request Prompt

5. CONCLUSION

In the field of multimedia interfaces, text-to-speech synthesis is a crucial research and application area. This paper evaluates key references to literature on endogenous speech signal variations and their significance in automatic text-to-speech recognition. If it produces natural speech, the text to conversion system can appear accurate and efficient to its users, and with a few tweaks, this system may be useful for blind people to interact with written documents. This paper included a step-by-step overview of the operation of a text-to-speech system (TTS). The program requests input data from the user in the form of a text file, which is then processed and translated into an mp3 file that can be played on any PC audio media player.

6. REFERENCES

[1] Smith R. (n.a). An Overview of the Tesseract OCR Engine", USA: Google Inc

[2] Teddy Surya Gunawan, Rashida Husain, Mira Kartiwi. (2017). Development of language identification system using MFCC and vector quantization, IEEE 4th International Conference on Smart Instrumentation, Measurement and Application (ICSIMA), pp.1-4.

[3] Rusli A. T., Ahmad M. I., Ilyas M. Z. (2018). Improving speaker verification using MFCC order, International

Conference on Robotics, Automation and Sciences (ICORAS), pp.1-4, 2016.

[4] Ashwin Nair Anil Kumar, Senthil Arumugam Muthukumaraswamy. (2017). Text dependent voice recognition system using MFCC and VQ for security applications, International conference of Electronics, Communication and Aerospace Technology (ICECA), Volume 2, pp.130-136.

[5] Mohsen Sadeghi, Hossein Marvi. (2017). OptimalMFCCFeaturesExtraction by Differential Evolution Algorithm for Speaker Recognition, 3rd Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS), pp.169-173.

[6] Mouaz Bezoui, Abdelmajid Elmoutaouakkil, Abderrahim Benihssane. (2016). Feature extraction of some Quranic recitation using Mel-Frequency Cepstral Coefficients (MFCC), 5th International Conference on Multimedia Computing and Systems (ICMCS), pp.127-131.

[7] Arpita Gupta and Akshay Joshi. (2018). Speech Recognitionusing Artificial NeuralNetwork, IEEE.

[8] Manjutha M, Gracy J, Subashini P, Krishnaveni M. (2017). Automated Speech Recognition System – A Literature Review",IJETA-V4I2P9.

[9] Ying Zhang, Mohammad Pezeshki, Phil´emonBrakel, Saizheng Zhang, C´esar Laurent Yoshua Bengio1, Aaron Courville. (2017). TowardsEnd-to-End Speech Recognition with Deep Convolutional Neural Networks, IEEE.

[10] Aditya Amberkar, Gaurav Deshmukh, ParikshitAwasarmol, Piyush Dave, "Speech Recognition using RecurrentNeural Networks, IEEE.

[11] JiPibil, Anna Pibilov, JindichMatouek. (2016). Comparison of one and two-level architecture of the GMM-based speaker age classifier", 39th International Conference on Telecommunications and Signal Processing (TSP), pp.299- 302.

[12] Rania Chakroun, Leila BeltafaZouari, MondherFrikha, Ahmed Ben Hamida. (2016). Improving text-independent speaker recognition with GMM, 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), pp.693-696.

[13] Suhas R. Mache, Manasi R. Baheti, Namrata C. Mahender. (2015). Review on Text-To-Speech Synthesizer, International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 8, August.

[14] Wenyong Lin. (2015). An improved GMM-based clustering algorithm for efficient speaker identification, 4th International Conference on Computer Science and Network Technology (ICCSNT), Volume 1, pp.1490-1493.

[15] Ledisi G. Kabari, Marcus B. Chigoziri. (2019). Speech Recognition Using MATLAB and Cross-Correlation Technique. *EJERS, European Journal of Engineering Research and Science Vol. 4, No. 8.*