iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.52906

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Autotuned Voice Cloning Enabling Multilingualism

Prof. Sakshi Shejole[1], Piyush Jaiswal[2], Neha Karmal[3], Vivek Patil[4], Samnan Shaikh[5]
[1, 2, 3, 4, 5]*Alard College of Engineering & Management (ALARD Knowledge Park, Survey No. 50, Marunji, Near Rajiv Gandhi IT Park, Hinjewadi, Pune-411057) Approved by AICTE Recognized by DTE NAAC Accredited. Affiliated to SPPU (Pune University)*

Abstract: ***This article describes a neural network-based text-to-speech (TTS) synthesis system that can generate spoken audio in a variety of speaker voices. We show that the proposed model can convert natural-language text-to-speech into a target language, and synthesize and translate natural text-to-speech. We quantify the importance of trained voice modules to obtain the best generalization performance. Finally, using randomly selected speaker embeddings, we show that speech can be synthesized with new speaker voices used in training and that the model learned high-quality speaker representations. We have also introduced a multilingual system and auto-tuner that allows you to translate regular text into another language, which makes multilingualization possible for various applications.***
Keywords: (Text to speech, Speech Synthesizer, Voice Cloning, Auto-tuner, Multilinguism) …

I. INTRODUCTION

Voice cloning uses a computer to generate voice from a real person and uses a neural network to clone that person's unique voice. This project uses a TTS system trained on a dataset consisting of text and speech. This allows the system to learn letters, words, and sentence sounds (such as waveforms). However, the resulting audio is the same as that represented by the training dataset. This means that the TTS system must be trained on the target speech to generate a specific speech. The text is then converted to normal speech. Synthetic speech can be generated by concatenating recorded speech segments. Additionally, synthesizers can combine speech models and other features of the human voice to create a fully "synthesized" speech output.

A. Voice Cloning

Voice cloning uses a computer to generate voice from a real person and uses a neural network to clone that person's voice. This model consists of an encoder and decoder and uses a vocoder to convert text to speech. After receiving the text data, the model recognizes the endpoints and evaluates the speech according to the condition whether the speech is clearly recognized. Also used auto tuner to clear the pitch and smooth the voice.

It currently consists of over 60 languages. According to the paper, modern multilingual text-to-speech systems require large amounts of data to train or process just a few languages, but deep learning techniques enable this model to train on small amounts of data and achieve high performance synthetic and stable voice cloning between multiple languages (English, German, French, Chinese, Russian).

B. Tortoise (Text-To-Speech) Synthesis

The goal of this paper is to make a TTS system that can induce natural speech for a variety of speakers in a data-effective manner. Speech synthesis is a technology that allows a computer to convert written text into speech via a microphone or telephone.

Tortoise is a text-to-speech synthesis system which describes a system, which produces synthetic speech. The program is organized on priority basis as followed as:

1) Powerful multi-voice functionality.
2) Very realistic prosody and intonation.

C. Auto Tuner

Auto-Tune uses a proprietary device to measure and alter the pitch of vocal and instrumental music recordings and performances. Training data consists of performance pairs that are identical except for pitch. Such pairs are needed for model training, but are difficult to find naturally.

Therefore, we construct the input signal by detuning high-quality vocal performances and synthesize the input signal by training a model to predict shifts that restore the original pitch.

©IJRASET: All Rights are Reserved

II. WORKING

Architecture diagrams create visual representations of software system components. In software systems, the term architecture refers to various functions, their implementation, and their interactions. It shows the general structure of a software system and the relationships, limits, and boundaries between individual elements.

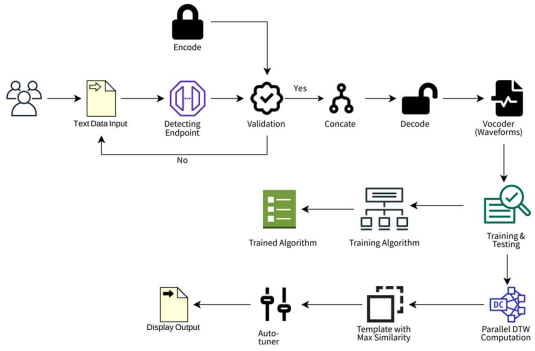


Diagram - 2.1: Architecture Diagram

A flowchart is a graphical representation of a step-by-step operation and action process with support for selection, iteration, and concurrency.

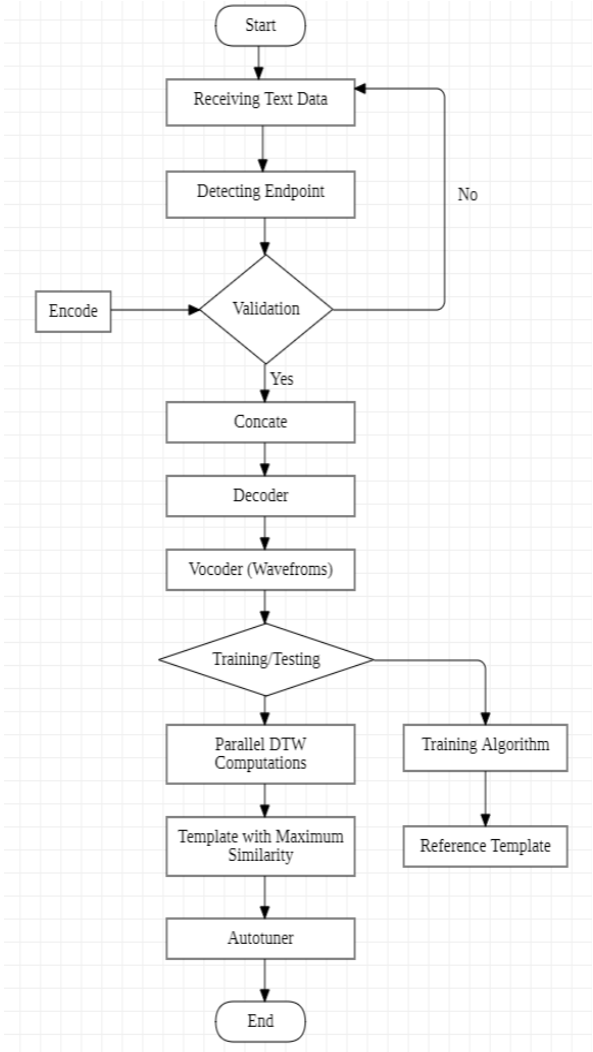


Diagram - 2.2: Flow Chart

III. OUTPUT

Installing various packages:

1) *Pytorch:* PyTorch is a Python package that provides two high-level features: Tensor computation (like NumPy) with strong GPU acceleration.
2) *Tortoise:* Tortoise is primarily an auto-regressive decoder model combined with a diffusion model. Its development prioritized realistic speech intonation and rhythm as well as multi-voice capabilities.
3) *Googletrans:* Googletrans is a free and unlimited python library that implements Google Translate API. It uses the Google Translate Ajax API to call methods such as Detect and Translate.

For successful installation of packages, a valid GitHub link should be provided with all the models. For installation of packages, we need GPU so make sure we have connected GPU during the runtime.

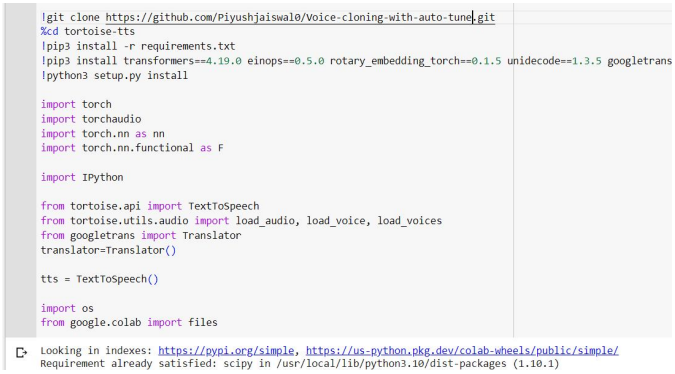


Diagram - 3.1: Install Packages

User interface: We have created a GUI for interaction with users which enables users to interact with the model easily. Here we should provide information regarding the input and required output.

Here are some inputs which should be filled by User:

a) *Sentence:* - Users should enter the sentence which they wanted to hear.
b) *Languages:* - Users should select the language in which they want to translate the given sentence
c) *Preset:* - User should select the audio quality. The default quality is “High quality”. Here we are using high quality natural voice as an output.
d) *Custom voice name:* - User should provide a unique name in which his/her voice model will be saved for testing.

Users should upload voice samples of the person in which they wanted to clone their voice. It should be noted that the user should upload 10 voice samples for better cloning of voice of 4-5 seconds each.

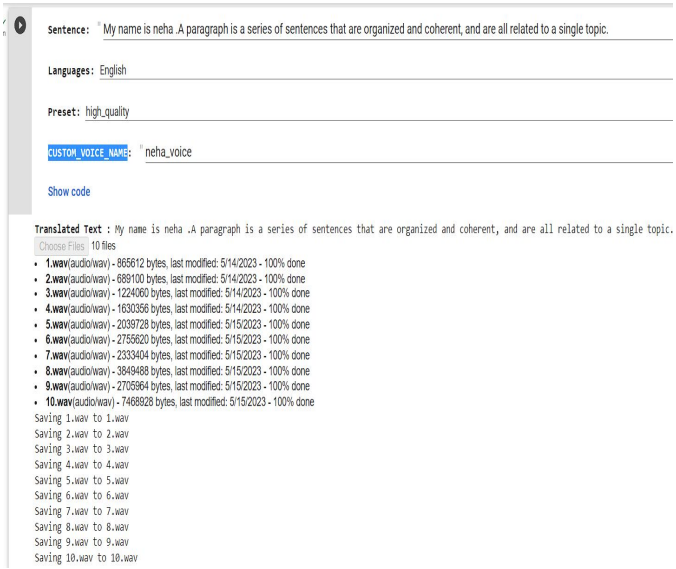


Diagram - 3.2: Graphical User Interface (GUI)

Finally, we will get an output as a cloned voice of the given sampled voice in a ".wav" file format which can be downloaded for future use.

```
# Generate speech
voice_samples, conditioning_latents = load_voice(CUSTOM_VOICE_NAME)
gen = tts.tts_with_preset(Text.text, voice_samples=voice_samples, conditioning_late
                          preset=Preset)
torchaudio.save(f'generated-{CUSTOM_VOICE_NAME}.wav', gen.squeeze(0).cpu(), 24000)
IPython.display.Audio(f'generated-{CUSTOM_VOICE_NAME}.wav')

/content/tortoise-tts/tortoise/utils/audio.py:14: WavFileWarning: Chunk (non-data)
  sampling_rate, data = read(full_path)
Generating autoregressive samples..
100%|██████████| 16/16 [03:17<00:00, 12.37s/it]
Computing best candidates using CLVP and CVVP
  0%|          | 0/16 [00:00<?, ?it/s]/usr/local/lib/python3.10/dist-packages/torch
  warnings.warn("None of the inputs have requires_grad=True. Gradients will be None
100%|██████████| 16/16 [00:32<00:00,  2.02s/it]
Transforming autoregressive outputs into audio..
100%|██████████| 400/400 [04:23<00:00,  1.52it/s]
▶ 0:00 / 0:14
```

Diagram - 3.3: Audio Output

IV. ACCURACY AND PRECISION

1) From the above research we noted that it will give higher accuracy if the training of the given samples are in the same language.

2) Hence we can increase the accuracy rate for better performance and good output for various applications in the specific domain.

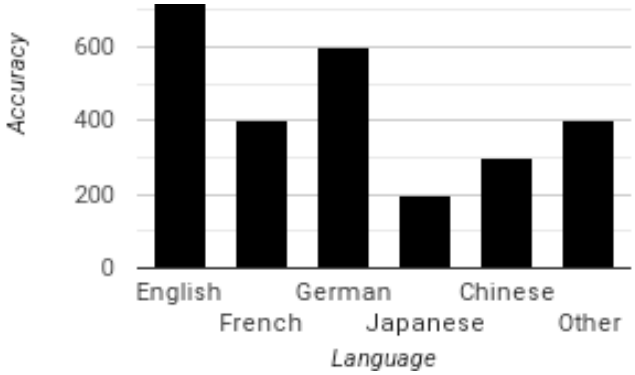


Diagram - 4.1: Accuracy Diagram

V. CONCLUSION

In this research paper, we have successfully studied Auto-tuned voice cloning which enables Multilingualism. In future, we are planning to use this model in Google Maps, Transportation services for creating a familiar voice to sound very natural and able to understand instructions fast and easily.

VI. ACKNOWLEDGEMENT

This paper was supported by Alard College of Engineering & Management, Pune 411057. We are very thankful to all those who have provided us valuable guidance towards the completion of this Seminar Report on "Autotuned voice cloning enabling multilingualism" as part of the syllabus of our course. We express our sincere gratitude towards the cooperative department who has provided us with valuable assistance and requirements for the system development. We are very grateful and want to express our thanks to Prof. Sakshi Shejole for guiding us in the right manner, correcting our doubts by giving us their time whenever we require, and providing their knowledge and experience in making this project.

REFERENCES

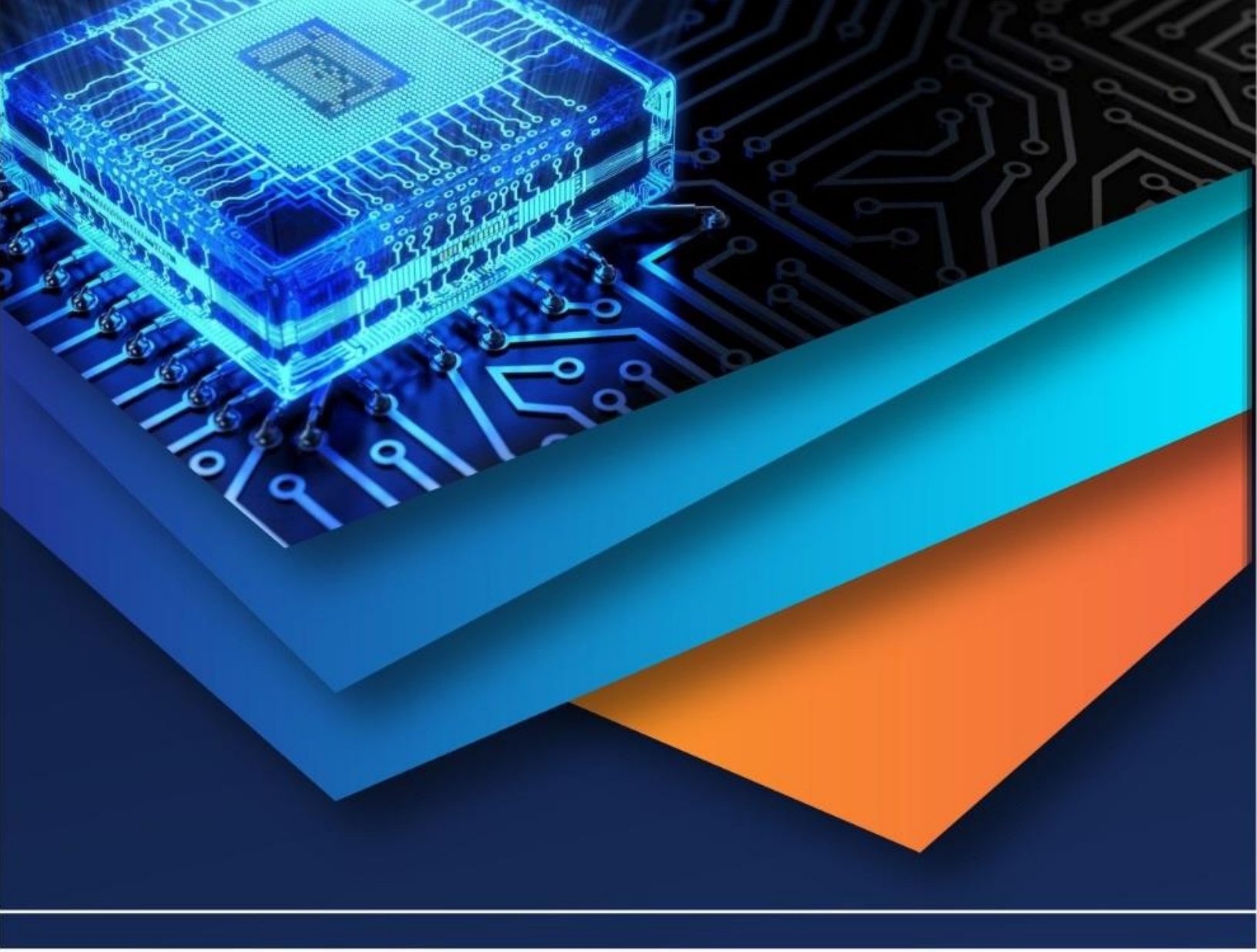
[1] Github links:https://github.com/jnordberg/tortoise-tts, https://github.com/CorentinJ/Real-Time-Voice-Cloning .

[2] Youtubelinks:https://www.youtube.com/watch?v=_iVu1oE8WGs,https://www.youtube.com/watch?v=O_hYhToKoA&t=31s,https://www.youtube.com/watch?v=Ewr7fpHiRvE&t=858s

[3] Dataset :https://drive.google.com/file/d/10q7nLKq9vmlGnHwWhBLPEnDVEQ849Kyo/view?usp=share_link,https://drive.google.com/drive/folders/17Y0WFC2chP2-3uM7nDUu5r3PUvdHX71k?usp=share_link

[4] Jiwon Seong and WooKey Lee, Suan Lee, "Multilingual Speech Synthesis for Voice Cloning" 2021 IEEE International Conference on Big Data and Smart Computing (BigComp) | 978-1-7281-8924-6/20/$31.00 ©2021 IEEE| DOI: 10.1109/BigComp51126.2021.00067

[5] Sanna Wager1 , George Tzanetakis2,3 , Cheng-i Wang3 , Minje Kim1 "DEEP AUTOTUNER: A PITCH CORRECTING NETWORK FOR SINGING PERFORMANCES"
[6] Nal Kalchbrenner * 1 Erich Elsen * 2 Karen Simonyan 1 Seb Noury 1 Norman Casagrande 1 Edward Lockhart 1 Florian Stimberg 1 Aaron van den Oord ¨ 1 Sander Dieleman 1 Koray Kavukcuoglu "Efficient Neural Audio Synthesis"
[7] Li Zhao , Li Zhao "Research on Voice Cloning with a Few Samples" 2020 International Conference on Computer Network, Electronic and Automation (ICCNEA)
[8] Ye Jia∗ Yu Zhang∗ Ron J. Weiss∗ Quan Wang Jonathan Shen Fei Ren Zhifeng Chen Patrick Nguyen Ruoming Pang Ignacio Lopez Moreno Yonghui Wu "Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis" arXiv:1806.04558v4 [cs.CL] 2 Jan 2019
[9] Qicong Xie1 , Xiaohai Tian2 , Guanghou Liu1 , Kun Song1 , Lei Xie1∗ , Zhiyong Wu3 , Hai Li4 , Song Shi4 , Haizhou Li2,5 , Fen Hong6 , Hui Bu7 , Xin Xu "THE MULTI-SPEAKER MULTI-STYLE VOICE CLONING CHALLENGE 2021"
[10] Li Wan Quan Wang Alan Papir Ignacio Lopez Moreno "GENERALIZED END-TO-END LOSS FOR SPEAKER VERIFICATION" arXiv:1710.10467v5 [eess.AS] 9 Nov 2020
[11] Yuxuan Wang∗ , RJ Skerry-Ryan∗ , Daisy Stanton, Yonghui Wu, Ron J. Weiss† , Navdeep Jaitly, Zongheng
[12] Yang, Ying Xiao∗ , Zhifeng Chen, Samy Bengio† , Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, Rif A. Saurous "TACOTRON: TOWARDS END-TO-END SPEECH SYNTHESIS": 6 Apr 2017
[13] Nwakanma Ifeanyi1 , Oluigbo Ikenna2 and Okpala Izunna3 "Text – To – Speech Synthesis (TTS)" IJRIT International Journal of Research in Information Technology.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)