Autotuned voice cloning enabling multilingualism

Prof. Priyadarshani Doke[1], Piyush Jaiswal[2], Neha Karmal[3], Vivek Patil[4], Samnan Shaikh[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, including those not seen during training. 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*

Key Words: (Text to speech, Speech Synthesizer, Voice Cloning, Auto-tuner, Multilinguism) …

1. Introduction

Voice cloning is the process in which one uses a computer to generate the speech of a real individual, creating a clone of their specific, unique voice using neural networks. A text-to-speech (TTS) system simply converts text to speech. In this project we are using TTS systems which are trained with datasets composed of texts and audio, thus, the system learns the sound (e.g., the waveform) of letters, words, and sentences. However, the resulting voice is the same as the one presented in the training dataset, which means that to produce a specific voice the TTS system needs to be trained with the target voice. Text is normal voice. Synthetic speech can be generated by concatenating recorded speech segments. In addition, synthesizers can combine voice models and other features of the human voice to produce a fully "synthesized" speech output.

1.1 Voice Cloning

Voice cloning is the process in which one uses a computer to generate the speech of a real individual, creating a clone of their specific, unique voice using neural networks. This model is composed of an encoder, and decoder and converts the text into audio using a vocoder. After receiving the text data the model detects the endpoint and evaluates the voice according to the condition that the voice is detected clearly or not. We are also using an auto-tuner for altering the tone, and pitch and smoothening the voice. At this time, composed of 60+ languages. According to the paper, the latest multi-lingual text-to-speech systems require a large amount of data for training or handling only two to three languages, but in this model, training with a small amount of data through deep learning technique enabled the high performance of synthetic sounds and stable voice-cloning between multiple languages (English, French, Chinese and Russian).

1.2 TTS

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. As an arising technology, not all inventors are familiar with speech technology. We specifically address a zero-shot literacy setting, where many seconds of un-transcribed reference audio from a target speaker is used to synthesize new speech in that speaker's voice, without streamlining any model parameters. Still, it's also important to note the eventuality for abuse of this technology, for illustration impersonating someone's voice without their concurrence. To address safety enterprises harmonious with principles similar, we corroborate that voices generated by the proposed model can fluently be distinguished from real voices.

1.3 Text-To-Speech Synthesis

A speech synthesis system is by description a system, which produces synthetic speech. It's implicitly clear, that this involves some kind of input. What isn't clear is the type of this input. However, which doesn't contain fresh phonetic and/ or phonological information the system may be called a Text-To-Speech (TTS) system, If the input is plain text. As shown, the conflation starts from text input. currently this may be plain text or pronounced-up text e.g., HTML or commodity analogous like JSML (Java Synthesis Mark- up Language).

1.4 Auto Tuner

Auto-Tune uses a proprietary device to measure and alter the pitch of vocal and instrumental music recordings and performances. The training data consists of performance pairs that are identical except for pitch. Such pairs are necessary for model training, but difficult to find naturally. Therefore, we construct input signals by detuning high-quality vocal performances and synthesize them by training a model to predict shifts that restore the original pitch.

3. Objectives of the study

It aims to induce synthetic voices veritably analogous to an original voice. Grounded on deep- literacy ways, this technology takes advantage of a set of audios of the original voice in order to train a model able of generating new audios that sound a suchlike

The specific objects are

1. To enable the deaf and dumb to communicate and contribute to the growth of an association through synthesized voice.

2. To enable the eyeless and senior people enjoy a stoner-friendly computer interface.

3. To produce ultramodern technology appreciation and mindfulness by computer drivers.

4. To apply an insulated whole word speech synthesizer that can convert text and responding with speech.

5. To measure and alter pitch in oral and necessary music recording and performances.

4. Scope of the study

Use-case diagrams describe the high-level functions and scope of a system. These diagrams also identify the interactions between the system and its actors. A Use case diagram outlines how external entities i.e. user interact with an internal software system.

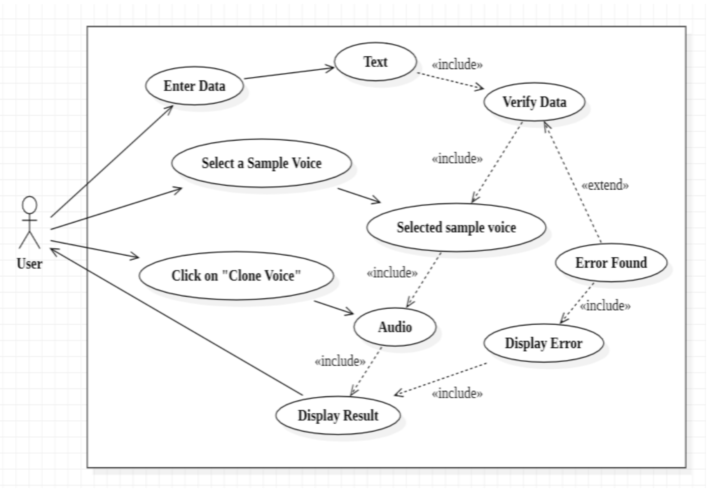

Diagram -1: Use Case Diagram

A state diagram consists of states, transitions, events, and activities. It describes the different states that an object moves through or provide an abstract description of the behavior of a system.

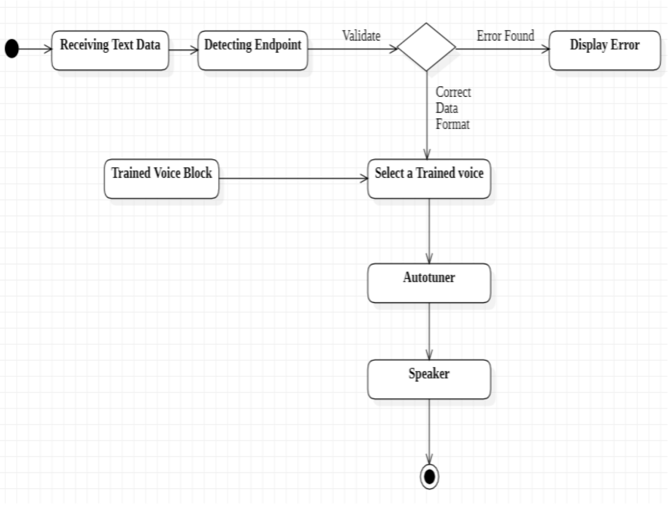

Diagram -2: State Diagram

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency.

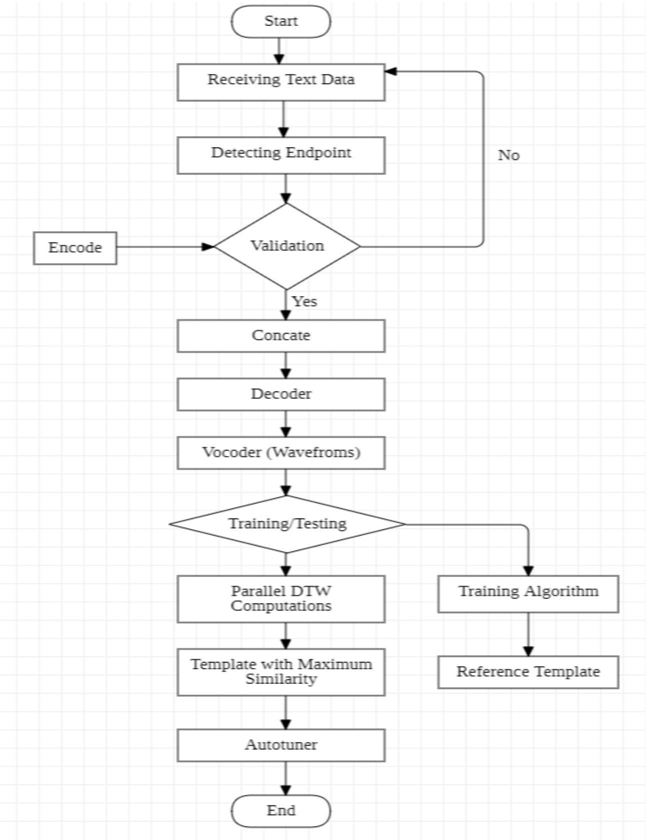

Diagram -3: Activity Diagram

3. CONCLUSION

In this research paper, we have successfully studied about 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.

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. Priyadarshini Doke for guiding us in the right manner, correcting our doubts by giving us their time whenever we required, and providing their knowledge and experience in making this project.

REFERENCES

[1] 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

[2] Sanna Wager1 , George Tzanetakis2,3 , Cheng-i Wang3 , Minje Kim1 "DEEP AUTOTUNER: A PITCH CORRECTING NETWORK FOR SINGING PERFORMANCES"

[3] 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"

[4] Li Zhao , Li Zhao "Research on Voice Cloning with a Few Samples" 2020 International Conference on Computer Network, Electronic and Automation (ICCNEA)

[5] 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

[6] 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"

[7] 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

[8] Yuxuan Wang∗ , RJ Skerry-Ryan∗ , Daisy Stanton, Yonghui Wu, Ron J. Weiss† , Navdeep Jaitly, Zongheng 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

[9] Nwakanma Ifeanyi1 , Oluigbo Ikenna2 and Okpala Izunna3 "Text – To – Speech Synthesis (TTS)" IJRIT International Journal of Research in Information Technology,