| PROJECT REQUIREMENTS SPECIFICATION  MultiSpeaker and Multilingual zero-shot Voice Cloning and Voice Conversion  UE19CS390A – Project Phase – 1  *submitted by*  **1.Anup Omkar - PES2UG1CS051**  **2.Kedarnath K Bhat - PES2UG19CS180**  **3.Nikitha M K - PES2UG19CS261**  **4. P Aftab Hussain - PES2UG19CS271**  Under the guidance of  **Prof.Ruby Dinakar**   | PES University | | --- |   **January - May 2022**  **DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**  FACULTY OF ENGINEERING  **PES UNIVERSITY**  (Established under Karnataka Act No. 16 of 2013)  Electronic City, Hosur Road, Bengaluru – 560 100, Karnataka, India |
| --- | --- |
|  |

TABLE OF CONTENTS

| 1. Introduction | 3 |
| --- | --- |
| 1.1 Project Scope and Motivation | 3 |
| 1. Literature Survey or Existing System | 3 |
| 1. Product Perspective | 3 |
| 3.1 Product Features | 3 |
| 3.2 User Classes and Characteristics | 3 |
| 3.3 Operating Environment | 3 |
| 3.4 General Constraints, Assumptions and Dependencies | 3 |
| 3.5 Risks | 4 |
| 1. Functional Requirements | 4 |
| 1. External Interface Requirements | 4 |
| 5.1 User Interfaces | 4 |
| 5.2 Hardware Requirements | 4 |
| 5.3 Communication Interfaces | 5 |
| 1. Non-Functional Requirements | 5 |
| 6.1 Performance Requirements | 5 |
| 6.2 Safety Requirements | 5 |
| 6.3 Security Requirements | 5 |
| 1. Other Requirements | 5 |
| Appendix A: Definitions, Acronyms, and Abbreviations | 5 |
| Appendix B: References | 5 |

# Introduction

The main objective of this document is to outline the requirements for the ‘Multispeaker and Multilingual Zero-shot Voice Cloning & Voice Conversion’. This document contains all of the software requirement specifics. It contains a general description of the types of users who will be using our application, how it is going to work, and what technologies we are using to make it work. We will also outline and describe specific components of the project.

# Project Scope

The main agenda of the project is “ voice conversion “ and “ voice cloning” by listening to any target speaker's voice in a specified language in order to get the desired result i.e either cloned or converted voice.

The goal of the project is to focus on building a model which focuses on a zero-shot learning approach.

# Literature Survey or Existing System

[Provide a detailed description of existing approaches, results, advantages, limitations of each paper.]

**1.Expressive voice conversion: A joint framework for speaker identity and emotional style transfer.**

* This paper proposed a GAN based framework to learn a many-to-many mapping across different speakers, that takes into account speaker-dependent emotional style without the need for parallel data. To this end, they condition the generator on emotional style encoding derived from a pre-trained speech emotion recognition (SER) model.
* StarGan,SER.
* Converts the voice only in a single language, can be extended to multiple languages

**2.GAZEV: GAN-Based Zero-Shot Voice Conversion over Non-parallel Speech Corpus.**

* This paper is a GAN-based zero-shot voice conversion solution, called GAZEV, which targets to support unseen speakers on both source and target utterances. This paper adopts speaker embedding loss on top of the GAN framework, as well as adaptive instance normalization strategy, in order to address the limitations of speaker identity transfer in existing solutions.
* GAZEV
* The further improvements can be made by adopting a better embedding module with more speech corpus data. Finding the best balance between the classifiers complexity and generative model complexity remains unclear.

**3.Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis.**

* This Paper talks about a neural network-based system for text-to-speech (TTS) synthesis that is able to generate speech audio in the voice of different speakers, including those unseen during training. This paper quantifies the importance of training the speaker encoder on a large and diverse speaker set in order to obtain the best generalization performance.
* VCTK dataset, LibriSpeech.
* The proposed model does not attain human-level naturalness, despite the use of a WaveNet vocoder. An additional limitation lies in the model’s inability to transfer accents. Finally, it is noted that the model is also not able to completely isolate the speaker voice from the prosody of the reference audio.

**4.Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning.**

* This paper presents a multispeaker, multilingual text-to-speech (TTS) synthesis model based on Tacotron that is able to produce high quality speech in multiple languages. Moreover, the model is able to transfer voices across languages, e.g. synthesize fluent Spanish speech using an English speaker’s voice, without training on any bilingual or parallel examples.
* Tacotron 2 neural TTS.
* Although, the model proposed performs well in intra language cloning but it performs poorly on interlanguage cloning.

**5.Neural Voice Cloning with a Few Samples.**

* For voice cloning the speaker characteristics are extracted for an unseen speaker from a set of cloning audios and generate an audio given any text for that speaker. The two performance metrics for the generated audio are speech naturalness and speaker similarity. Speaker adaptation is the idea to ﬁne-tune a trained multi-speaker model for an unseen speaker using a few audio-text pairs.

Speaker encoding is a method to directly estimate the speaker embedding from audio samples of an unseen speaker. It does not require any ﬁne-tuning during voice cloning. Thus, the same model can be used for all unseen speakers.

* VCTK Dataset
* The naturalness can even further be improved with ﬁne-tuning. Similarity scores slightly improve with higher sample counts for speaker encoding, and match the scores for speaker embedding adaptation. Thus, the proposed techniques could be improved with better multi-speaker models. (like using WaveNet vocoder instead of Grifﬁn- Lim). The drawbacks could be observed by training the model using a speech recognition dataset with low-quality audios and limited speaker diversity.

**6.YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for everyone.**

* As a vocoder the HiFi-GAN version 1 is used. For efﬁcient end2end training, the TTS model is connected with the vocoder using a variational autoencoder. In parallel with the ZS-TTS, multilingual TTS has also evolved aiming at learning models for multiple languages at the same time. Some of these models are particularly interesting as they allow for code-switching. YourTTS is built upon VITS, but includes several novel modiﬁcations for zero-shot multi-speaker and multilingual training. For multilingual training, 4-dimensional trainable language embeddings are concatenated into the embeddings of each input character. 3 languages were investigated, using one dataset per language to train the model. English: VCTK dataset, which contains 44 hours of speech. Portuguese: TTS-Portuguese Corpus, and French: fr\_FR set of the M-AILABS dataset, consists of 2 female and 3 male speakers.
* VAE's Hifi-GAN's
* Mispronunciations occur for some words, especially in Portuguese. They do not use phonetic transcriptions, making our model more prone to such problems.k improvements to the duration predictor of the YourTTS model as well as training in more languages.

**7.Zero-Shot Text-to-Speech for Text-Based Insertion in Audio Narration.**

* Given a piece of speech and its transcript text, text-based speech editing aims to generate speech that can be seamlessly inserted into the given speech by editing the transcript. The paper manages to perform accurate zero-shot duration prediction for the inserted text. The predicted duration is used to regulate both text embedding and speech embedding. Then, based on the aligned cross-modality input, they directly generate the mel-spectrogram of the edited speech with a transformer-based decoder.
* CNN, Short time Fourier Transforms, Griffin-Lim algorithm in Vocoder.
* Although the model performs well, it is trained only on English language and there is no multilingual or inter language Conversion.

**8.Expressive Neural Voice Cloning**

* Voice cloning is the task of learning to synthesize the voice of an unseen speaker from a few samples. While current voice cloning methods achieve promising results in Text-to-Speech synthesis for a new voice, these approaches lack the ability to control the expressiveness of synthesized audio. Paper proposes a controllable voice cloning method that allows fine grained control over various style aspects of the synthesized speech for an unseen speaker. It is achieved by explicitly conditioning the speech synthesis model on a speaker encoding, pitch contour and latent style tokens during training. These cloning tasks include style transfer from a reference speech, synthesizing speech directly from text, and fine-grained style control by manipulating the style conditioning variables during inference.
* Tacotron 2, GST model
* The model is trained only on English Corpus and hence can't clone the voice in other languages

# Product Perspective

This product is designed to run any system that is capable of running machine learning code snippets. We are looking to make a web app for the user so that he does not need to install all dependencies. Hence, the sole requirement for the user is a web browser (Safari, Firefox, or Internet Explorer) with an active internet connection.

# Product Features

The major features of the product:

1. Voice Cloning: The user should provide/input his voice for a few seconds to a couple of minutes. Then the text is given as input, which is read aloud by the model just like TTS, in the target speaker’s voice.
2. Voice Conversion: The user should provide/input his voice for a few seconds to a couple of minutes. Now the voice of one user is converted to the target speaker’s voice.

# User Classes and Characteristics

1. Input Voice: The voice is given as input to the model, which must be at least 5 seconds and the maximum limit can be of user choice.
2. Text characters: The text given as input determines the time for processing. If the user wants the output as fast as possible, he can give input in multiple text boxes and get output. But the main disadvantage is that he/she gets the voice in discrete form. But if the user gives the input in a single text box with multiple characters, then he gets the continuous voice but takes much time to process.
3. Voice Library: The User can even be able to select the voices from the list of voices given for voice conversion purposes.
4. Login System: To prevent the users/people to use the model in an abusive manner, we keep track of their usage by login system.

# Operating Environment

The System can be operated in all the major platforms with a Good GPU like Nvidia. if he wants to run the product by cloning the code, otherwise he just requires a browser with a good internet connection.

# General Constraints, Assumptions, and Dependencies

* Regulatory policies: The users can use the model in an abusive way i.e, they convert their voice to a person’s voice and use the voice in an illegal way
* Hardware limitations: Requires a lot of GPUs for training
* Parallel operations: During the training phase of the model, the model is trained in a distributed environment, so that model parameter can be tuned easily in less time to make the model more efficient.
* Criticality of application: While recording the voice, the user must be in a noise-free area so that the model predicts accurately.
* Safety and security considerations: To prevent the misuse of the product, the IP address of the user is taken and a login system is made, in which the user can log in only through his phone number only, to prevent the creation of fake accounts.

# Risks

1. Taking care of emotions and expressions while voice conversing in different languages.
2. The GPUs are required to train a model which costs more.
3. Dataset for Kannada, Hindi languages is scarce.
4. Make sure that the target output of the English Language matches with the accent (USA, UK, India) of the user’s country (if provided), Distributed training on cloud platforms to make efficient use of GPUs and train in less time.

# Functional Requirements

* **Input**: The input voice (i.e target voice) given by the user must be above 5 seconds and in Voice Cloning, text input given by the user must contain a minimum of 20 proper characters (i.e characters without punctuations) for correct generation of voice.
* **The sequence of operations**:
  + The user has to provide his / her location access and IP address access.
  + The captcha is provided, to ensure that the user is a human and not a bot.
  + Then, the user is allowed to log in, with his mobile number.
  + After Successful login, the user can record his voice (which is Target voice)
  + If the user wants to do Voice Cloning, he/she can provide a text in a text box with correct input as specified in the above point and select the language.
  + If the user wants to convert his / her voice to another person's voice, then select the voice from the available list of voices or provide the other person’s voice as input. If the user wants to convert his own voice in his own voice but with a different language, then he can do so by selecting the language specified.
  + The languages for cloning and conversion of voice include only English, Kannada & Hindi (Tentative).
* **Error handling and recovery:**
  + If the user didn’t allow the system to access his location then he/she will not be allowed to login into his / her account.
  + If the user stops the location access to his/ her account after his login, then the user will be logged out of his account automatically.
  + If the input provided is not valid, then the user is asked to give a valid input.
* Increased length of the reference speech significantly improves the similarity. Shorter reference utterances give slightly better naturalness.

# External Interface Requirements

# User Interfaces

* The UI is a simple web application.
* The Voice recording option is given in a web application or the user can even input the previously recorded audio file.
* The Input text box is provided or the text file can be uploaded for Voice Cloning purposes.
* The checkboxes or radio buttons are provided for selecting the options available according to the user needs.

# Hardware Requirements

Requires GPUs for model training if the user wants to use the source code and modify it for enhancements.

# Communication Interfaces

The microphone is required for the audio recording.

# Non-Functional Requirements

# Performance Requirement

The response time / the output voice provided by the system depends on the user input audio length and the text length provided. A good internet connection of at least 5 Mbps fetches the output within a minute.

# Safety Requirements

IP address & Location access must be given by the user to ensure that he do not use the feature for illegal purposes

# Security Requirements

A login authorisation system using only mobile number is done.

# Other Requirements

None as of now.

# Appendix A: Definitions, Acronyms, and Abbreviations

**Voice Cloning:** Voice cloning is when a computer program is used to generate a synthetic, adaptable copy of a person's voice.

**Voice Conversion**: Voice Conversion aims to modify an utterance from the source speaker to make it sound like the target speaker.

**GPU:** Graphics processing unit, a specialized processor originally designed to accelerate graphics rendering. GPUs can process many pieces of data simultaneously, making them useful for machine learning, video editing, and gaming applications.

# Appendix B: References