AI VOICE CLONING

A Software Requirement Specification report submitted for the
Award of Degree of

BACHELORS DEGREE
in
COMPUTER SCIENCE AND ENGINEERING
by

MUBEEN SHAIK
ID NO: B151743

PRASHANTH SIRUSALA
ID NO:B151796

April 24, 2021
Under the supervision of

Mr. Laxmi Narayana Garu
Assistant Professor at RGUKT Basar
(Duration : 02/2021 to 07/2021)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
Rajiv Gandhi University of Knowledge and Technologies
Basar,Nirmal,Telangana,INDIA.

July 2020-2021

Contents

Chapter 1

Introduction

The time when voice assistants were first introduced on smartphones. Although the capabilities of voice assistants were very limited back then. But, the concept of talking to a phone got everyone excited. One of the limitations of early voice assistants was that they used to sound very unnatural and robotic. But, as time progressed, both spoken and functional capabilities of these assistants have also evolved. Now, the accent, tone, pitch, etc. of voice agents sounds very natural. It all has become possible because of the advancements in the field of artificial intelligence and text-to-speech over the years.

Neural networks can now take just a few seconds of your speech and generate entirely new natural-sounding audio samples. What's more, these synthetic voices may soon be indistinguishable from the original audio samples. If you try to examine the several examples of voice cloning, it's easier to appreciate the breadth of what the technology can do including being able to switch the gender of the voice and alter the accents and styles of speech.

Slowly we are moving toward a voice-driven world. The consumption of audio content and voice-based automated services is on the rise. Many content creators are moving to platforms like SoundCloud and Amazon's audiobook service, Audible. It can also be sensed from the fact that tech giants like Google, Amazon, Samsung, Apple, etc, are investing heavily in their voice-based services, and very often they claim to be better than their counterparts. With these advancements, soon we will be able to customize the voice of various voice agents as we like.

1.1 Purpose

This Software Requirements Specification aims to mark out and show the design and architecture of AI Voice Cloning from different aspects. It describes how the system is structured in order to supply the demands mentioned in this document. It is intended to give a hand to the reader to understand the implementation phase. The main purpose of this project is try and eliminate the robotic tone when a voice is generated.

1.2 Intended Audience and Reading Suggestions

The main audience for this document is the design and the development team of the Smart Driver Assistant project. Development, testing, and design are all done by the same team. Further the discussion will provide all the internal, external, functional and also non-functional informations about "AI Voice Cloning".

1.3 Project Scope

In a bid to sound more like humans, artificial intelligence (AI) is all set to break new records, literally. A new technology, called 'Voice Cloning' is replacing the robotic tonality of virtual assistants with natural human voices. Voice cloning with artificial intelligence can master unique human voices to make chatbots, video clips, and other interactions more intuitive and engaging.

AI's underlying technologies, machine learning and deep learning have constantly demonstrated significant potential for text-to-speech (TTS) interactions, also called speech synthesis. The technology when coupled with speech recognition becomes the backbone for virtual assistants such as Siri, Alexa, and the likes. However, providers of chatbot development services still struggle at eliminating the robotic tonality associated with voice-controlled assistants.

With voice cloning, deep neural networks are moving a step closer to interactive, personalized, and highly intuitive human-chatbot interactions.

A recent research paper, Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis by Jia, Zhang, and others introduce an arguably earlier and more efficient way for voice cloning. The paper proposes a new technique, Speech Vector to TTS (SV2TTS) that generates near-similar speech audio using only a few seconds of a sample voice. Unlike highly expensive traditional training methods that required several hours of professionally recorded speech, SV2TTS can-

a) Clone voices without excessive training or retraining

b) Produce high-quality audio results, and

c) Synthesize natural speech from speakers unseen during the training.

Chapter 2

Overall Description

2.1 Product Perspective

Text to Speech Synthesis is a problem that has applications in a wide range of scenarios. They can be used to read out pdf's loud, help the visually impaired to interact with text, make chatbots more interactive etc. Historically, many systems were built to tackle this task using signal processing and deep learning approaches.In this article, this project explore a novel approach to synthesize speech from the text presented by Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno and Yonghui Wu, researchers at google in a paper published on 2nd January 2019.

MultiSpeaker Text to Speech synthesis refers to a system with the ability to generate speech in different users voices. Collecting data and training on it for each user can be a hassle with traditional TTS approaches.SV2TTS is a three-stage deep learning framework that allows to create a numerical representation of a voice from a few seconds of audio, and to use it to condition a text-to-speech model trained to generalize to new voices.

2.2 Product Functions

The goal of this work is to build a TTS system which can generate natural speech for a variety of speakers in a data efficient manner. We specifically address a zero-shot learning setting, where a few seconds of untranscribed

reference audio from a target speaker is used to synthesize new speech in that speaker's voice, without updating any model parameters. Such systems have accessibility applications, such as restoring the ability to communicate naturally to users who have lost their voice and are therefore unable to provide many new training examples. They could also enable new applications, such as transferring a voice across languages for more natural speech-to-speech translation, or generating realistic speech from text in low resource settings.

2.3 Characteristics

A new approach with three independent components is introduced to provide an efficient solution to the multi-speaker adaptation during speech synthesis. These components are deep learning models that are trained independently of each other. Let's understand what each of these components are doing.

1.Speaker Encoder

Each speaker's voice information is encoded in an embedding. This embedding is generated by a neural network trained using speaker verification loss. Speaker verification loss is calculated by trying to predict whether two utterances are from the same user or not.

Speaker Embeddings

The embeddings are supposed to have high similarity if and only if they are from the same user.

Note that this training need not have any information about the text that we are trying to vocalize. Moreover, once trained on a large corpus of unlabelled voices containing background noises and disturbances, the model develops an ability to learn crucial information regarding the speaker's voice characteristics. This enables us to generate embeddings for new users without having to change the network's parameters.It won't even require more than a few seconds of the target user's voice utterance of any text.

This makes the embeddings agnostic to the downstream task,allowing them to be trained independent of the synthesis models that follow.

2.Synthesizer

This component is the core model of Text-to-Speech Synthesis. It takes in the sequence of phonemes as inputs and generates a spectrogram of the corresponding text input. Phonemes are distinct units of a sound of words. Each word is decomposed into these phonemes and sequence input to the model is formed. This model also consumes Speaker encodings to support MultiSpeaker Voices. Following is the high-level overview of this model.

Speaker encoding is concatenated with each layer's output. This speaker encoding can be generated in the previous step completely unaware of the current phoneme sequence.In Fact these embedding can even be random samples from the distribution of these encodings. If we give random vectors this model will generate a synthetic voice that resembles human voices.

Training of this model is done by minimizing L2 loss of the generated spectrogram. Mel Spectrogram for targets is obtained by breaking down audio into time segments, calculating the frequency components and converting them into Mel Scale. Mel Scale is a fixed non-linear transformation of inputs. Mel Scale transforms the frequency scale into a human perceptual scale.

Reconstruction of audio from spectrograms is not as trivial as generating the spectrogram from audio samples. To generate audio we use the following vocoder network.

3.Vocoder

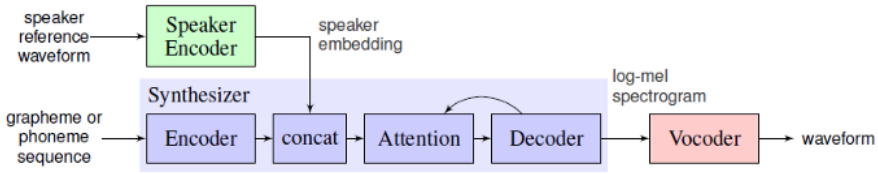
A sample by sample autoregressive WaveNet model is used to perform voice generation. This model takes Mel Spectrogram as input to generate time-domain waveforms. Following is the architecture of this WaveNet Model

The dilated convolution blocks used in the model are quite interesting. They enforce sequence data's causality by restricting the convolution to only look at values from previous time steps. But this narrows down the receptive fields of neurons resulting in the requirement of very high depth models. Dilation is a nice concept that skips over a few neurons in previous time steps to increase each neuron's range in deeper layers.

This model doesn't need a separate representation of the target speaker as the spectrogram contains all the information. Once the model is trained on a large enough corpus containing multiple speakers' voice, It becomes good at generating voices of unknown speakers.

Inference

Inference can be done on this model using zero shot transfer learning. We just need a few second speech sample of a new user and the model adapts to the speaker's characteristics and generates a speaker encoding which can be used along a target text to synthesize speech.



2.4 Operating Environment

The GUI will be operable in any Operating Environment - Mac, Windows, Linux etc.

2.5 Design and Implementation Constraints

1. Experimental support for low-memory GPUs (2gb) added for the synthesizer.
2. It's not recommended if you have enough VRAM.
3. Need to download pretrained models

Chapter 3

Literarture Survey

3.1 Introduction to domain

Digital cloning is an emerging technology, that involves deep-learning algorithms, which allows one to manipulate currently existing audio, photos, and videos that are hyper-realistic. One of the impacts of such technology is that hyper-realistic videos and photos makes it difficult for the human eye to distinguish what is real and what is fake. Furthermore, with various companies making such technologies available to the public, they can bring various benefits as well as potential legal and ethical concerns.

Voice cloning is a deep-learning algorithm that takes in voice recordings of an individual and is able to synthesize such a voice into one that is very similar to the original voice. Similar to deepfakes, there are numerous apps, such as Resemble AI, iSpeech, and CereVoice Me, that gives the public access to such technology. The algorithm simply needs at most a couple of minutes of audio recordings in order to produce a voice that is similar and it will also take in any text and will read it out loud. Although this application is still in the developmental stage, it is rapidly developing as big technology corporations, such as Google and Amazon are investing huge amounts of money for the development.

Some of the positive uses of voice cloning include the ability to synthesize millions of audiobooks without the use of human labor. Another include those who may have lost their voice can gain back a sense of individuality

by creating their own voice clone by inputting recordings of them speaking before they lost their voices. On the other hand, voice cloning is also susceptible to misuse. An example of this is voices of celebrities and public officials being cloned and the voice may say something to provoke conflict despite the actual person has no association to what their voice said.

3.2 Data Methodology

Based on the research paper. two public datasets for training the speech synthesis and vocoder networks. VCTK contains 44 hours of clean speech from 109 speakers, the majority of which have British accents are available. That can be downsampled the audio to 24 kHz, trimmed leading and trailing silence (reducing the median duration from 3.3 seconds to 1.8 seconds), and split into three subsets: train, validation (containing the same speakers as the train set) and test (containing 11 speakers held out from the train and validation sets).

LibriSpeech consists of the union of the two "clean" training sets, comprising 436 hours of speech from 1,172 speakers, sampled at 16 kHz. The majority of speech is US English, however since it is sourced from audio books, the tone and style of speech can differ significantly between utterances from the same speaker. They resegmented the data into shorter utterances by force aligning the audio to the transcript using an ASR model and breaking segments on silence, reducing the median duration from 14 to 5 seconds. As in the original dataset, there is no punctuation in transcripts. The speaker sets are completely disjoint among the train, validation, and test sets.

Many recordings in the LibriSpeech clean corpus contain noticeable environmental and stationary background noise. They preprocessed the target spectrogram where the background noise spectrum of an utterance was estimated as the 10th percentile of the energy in each frequency band across the full signal. This process was only used on the synthesis target; the original noisy speech was passed to the speaker encoder.

3.3 How different from existing systems

Synthesizing natural speech requires training on a large number of high quality speech-transcript pairs, and supporting many speakers usually uses tens of minutes of training data per speaker. Recording a large amount of high quality data for many speakers is impractical. Our approach is to decouple speaker modeling from speech synthesis by independently training a speaker-discriminative embedding network that captures the space of speaker characteristics and training a high quality TTS model on a smaller dataset conditioned on the representation learned by the first network. Decoupling the networks enables them to be trained on independent data, which reduces the need to obtain high quality multispeaker training data. We train the speaker embedding network on a speaker verification task to determine if two different utterances were spoken by the same speaker. In contrast to the subsequent TTS model, this network is trained on untranscribed speech containing reverberation and background noise from a large number of speakers.

Cloning a voice typically requires collecting hours of recorded speech to build a dataset then using the dataset to train a new voice model. But not anymore. This project introduces a remarkable Real-Time Voice Cloning Toolbox that enables anyone to clone a voice from as little as five seconds of sample audio.

Chapter 4

Software Requirement Specification

4.1 SDLC Methodology

Agile Methodology

The Agile software development methodology is one of the simplest and effective processes to turn a vision for a business need into software solutions. Agile is a term used to describe software development approaches that employ continual planning, learning, improvement, team collaboration, evolutionary development, and early delivery. It encourages flexible responses to change.

The agile software development emphasizes on four core values.

1. Individual and team interactions over processes and tools
2. Working software over comprehensive documentation
3. Customer collaboration over contract negotiation, here the customer i.e end-user will be the team itself.
4. Responding to change over following a plan

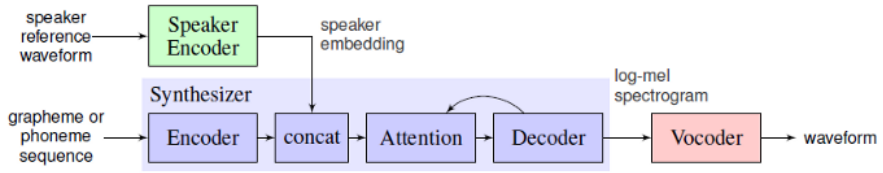
4.2 Agile Project Management Tool

JIRA

Jira Software is an agile project management tool that supports any agile methodology, be it scrum, kanban, or your own unique flavor. From agile boards, backlogs, roadmaps, reports, to integrations and add-ons you can plan, track, and manage all your agile software development projects from a single tool.

Create tasks for yourself and members of your team to work on, complete with its details, due dates, and reminders. Utilize subtasks to breakdown larger items of work. Allow others to watch the task to track its progress and be notified when it's completed. Create sub-tasks within the parent task to break down the unit of work into digestible pieces for various members of the team. View all tasks on the board to easily visualize each's status.

4.3 Architecture



This system consists of three independently trained components. This allows each component to be trained on independent data, reducing the requirement of high-quality multispeaker data. The individual components are:

Speaker Encoder Network

The speaker encoder network's job is to take audio of a given speaker as input, and encode the characteristics of their voice into a low dimensional vector embedding. It does not care about what the speaker is saying, all it cares about is how a speaker is saying something. The network is trained separately on the task of speaker verification, using a dataset of noisy speech from

thousands of speakers. The encodings are then used to condition the synthesis network on a reference speech signal from the desired target speaker.

Synthesis Network

It is a Seq2Seq neural network based on google's Tacotron 2 that generates a Mel spectrogram from the text, conditioned on the speaker embedding. The network is an extended version of Tacotron 2 that supports multiple speakers. The output is conditioned according to the voice of the speaker by concatenating their embedding with the synthesizer encoder output at each time step.

Vocoder Network

The system uses the WaveNet as a vocoder. It takes the Mel spectrograms generated by the synthesis network as input and autoregressively generate the time-domain audio waveforms as output. The synthesizer network is trained such that, it tries to capture all of the relevant detail needed for the high-quality synthesis of a variety of voices in the form of Mel spectrograms. This allows the vocoder to be constructed by simply training on data from many speakers.

Chapter 5

Requirements

5.1 Functional Requirements

This project is build using,

Hrdware

An NVIDIA graphics card that supports CUDA 10

Dataset

LibriSpeech

Back-End

1.Python x86-64 3.7

2. PyTorch for CUDA 10.0
This is done via Python's pip, you do not download anything manually
Instructions for this are provided further down

3. NVIDIA CUDA 10.0
This is because TensorFlow >=1.10.0, <=1.14 is required

Download the network install since you won't need all the components unless you have another need for CUDA
NVIDIA cuDNN for CUDA 10.0

4. Visual Studio Community 2014 or later
2019 is the latest, and yes you can use it

5. Pretrained models
These are necessary if you aren't training the models yourself

GUI

PyQt5

Python packages

umap-learn
visdom
librosa>=0.8.0
matplotlib>=3.3.0
numpy==1.19.3
numpy==1.19.4
scipy>=1.0.0
tqdm
sounddevice
SoundFile
Unidecode
inflect
PyQt5
multiprocess
numba
webrtcvad

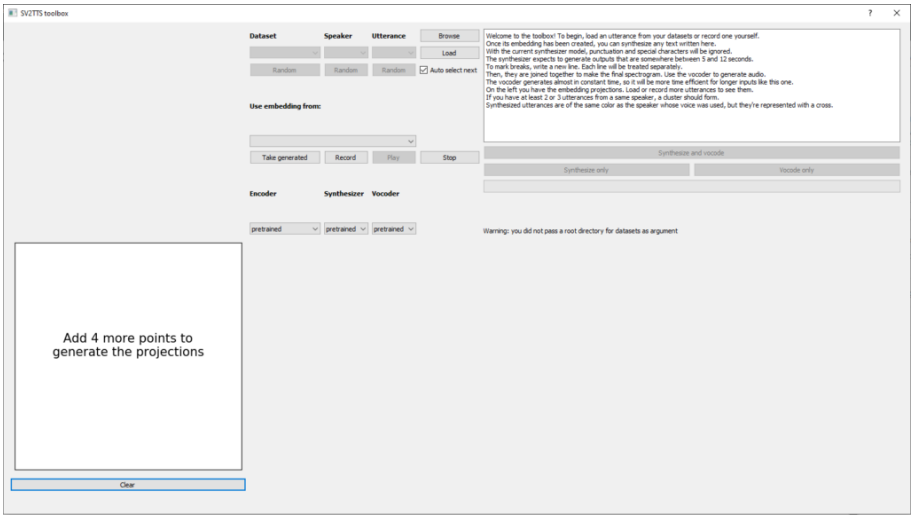
Audio Operations

FFMPEG

5.2 External Interfaces Requirements

PyQt5 Widgets

Widgets are basic building blocks of an application. PyQt5 has a wide range of various widgets, including buttons, check boxes, sliders, or list boxes.



Chapter 6

Other Nonfunctional Requirements

6.1 Safety Requirements

Although the concept of artificial voice is fascinating and has many benefits, we can't deny the fact that this technology is susceptible to misuse. In the past few years, we have seen how Deepfakes were being used to spread misinformation and to create questionable content.

As the voice cloning algorithms are getting better, it is becoming more and more difficult to discern what's real and what's not. Using voice cloning people can be fooled to act on something fake just because it sounds like it's coming from somebody real. For instance, it will become easier for scammers to perform phishing and spoofing attacks, things that people never uttered could be pushed on the internet in a planned manner for political gains, fake audio clips could also be used to create unrest in society, and the list goes on.

According to research, the human brain does not register significant differences between real and artificial voices. In fact, it is harder for our brains to distinguish fake voices than to detect fakes images. So, raising the awareness that this technology exists will be the first step toward safeguarding the listeners. Algorithms that can differentiate real voices from artificial voices should be developed alongside.

Owing to the ethics associated with this technology, many are skeptical if humans should even try creating such models. Some other researchers have refrained from sharing their models publicly.

Nevertheless, the future appears to be uncertain, as to how humanity will use this technology and what transpires from it, a dystopia or a utopia?

In order to address safety concerns consistent with principles such as above, we verify that voices generated by the proposed model can easily be distinguished from real voices.

6.2 Security Requirements

Voice cloning is also susceptible to misuse. An example of this is voices of celebrities and public officials being cloned and the voice may say something to provoke conflict despite the actual person has no association to what their voice said.

In recognition of the threat that voice cloning poses to privacy, civility, and democratic processes, the Institutions including FTC, U.S. Department of Justice and Defense Advanced Research Projects Agency (DARPA) have weighed in on various deepfake audio use cases and methods that might be used to combat them.

6.3 Software Quality Attributes

In the development phase testing is been continued. So that the quality of the software is been maintained and all the requirements are been fulfilled.UI test is done.

Chapter 7

Appendices

7.1 Appendix A: Glossary

AI - Artificial Intelligence
ASR - Automatic Speech Recognition
FTC - Federal Trade Comission
TTS - Text To Speech
VCTK - VCTK Corpus Dataset
GUI - Graphical User Interface
CUDA - Compute Unified Device Architecture
VRAM - Video Random Access Memory
SV2TTS - Speaker Verification to Text To Speech

7.2 Appendix C: To Be Determined List

Wave Net Research Papers
Synthesis Model
Speaker Verification Research Papers
SV2TTS Research Papers