Report On

Clone Voice AI

Submitted in partial fulfillment of the requirements of the Course project in Semester VIII of fourth year Artificial Intelligence and Data Science

by

Harshal Ahire (Roll No. 02)
Ganesh (Roll No. 18)
Sanket Das (Roll No. 15)

Supervisor
Dr. Tatwadarshi P. N.

University of Mumbai

Vidyavardhini's College of Engineering & Technology

Department of Artificial Intelligence and Data Science

(2024-25)

Vidyavardhini's College of Engineering & Technology Department of Artificial Intelligence and Data Science

CERTIFICATE

This is to certify that the project entitled “Clone Voice AI” is a bonafide work of “Harshal Ahire (Roll No. 02), Ganesh (Roll No. 18), Sanket Das (Roll No. 15)” submitted to the University of Mumbai in partial fulfillment of the requirement for the Course project in Semester VIII of fourth year Artificial Intelligence and Data Science engineering.

Supervisor

Dr. Tatwadarshi P. N.

Dr. Tatwadarshi P. N.
Head of Department

Table of Contents

Chapter 1: Introduction

1.1 Introduction

Deep learning models have become increasingly popular in computational machine learning, such as Text-to-speech (TTS). Deep models that create more natural-sounding speech than the conventional concatenative methods began emerging in 2016. Research has focused on making these models more effective, sound more natural, or training them in an end-to-end fashion. Inference on GPU has come from being hundreds of times slower than real-time on a mobile CPU. She et al. (2017) showed near-human naturalness as to the quality of the speech produced. Low- dimensional embedding is derived from a speaker encoder model which takes reference speech as input. This approach is more data-efficient than training a separate TTS model for each speaker, as well as faster and less computationally expensive orders of magnitude. There is a broad disparity between the length of reference speech necessary to clone a voice among the various methods, ranging from half an hour per speaker to just a few seconds.

1.2 Problem Statement & Objectives

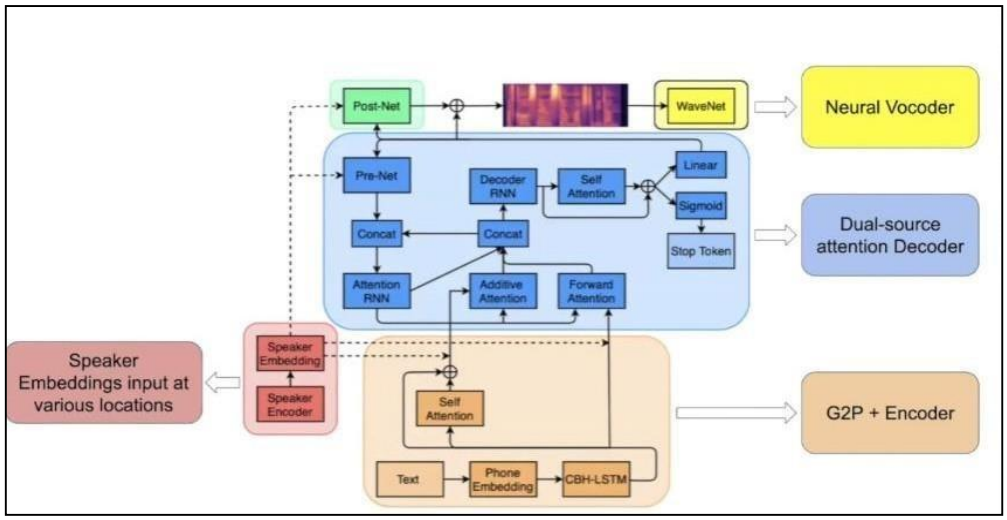
The resulting framework must be able to operate in a zero-shot setting, that is, for speakers unseen during training. It should incorporate a speaker's voice with only a few minutes of reference speech. In addition, we integrate a model based on the framework to make it run in real-time, i.e. to generate speech in a time shorter or equal to the duration of the produced speech.

Chapter 2: Proposed System

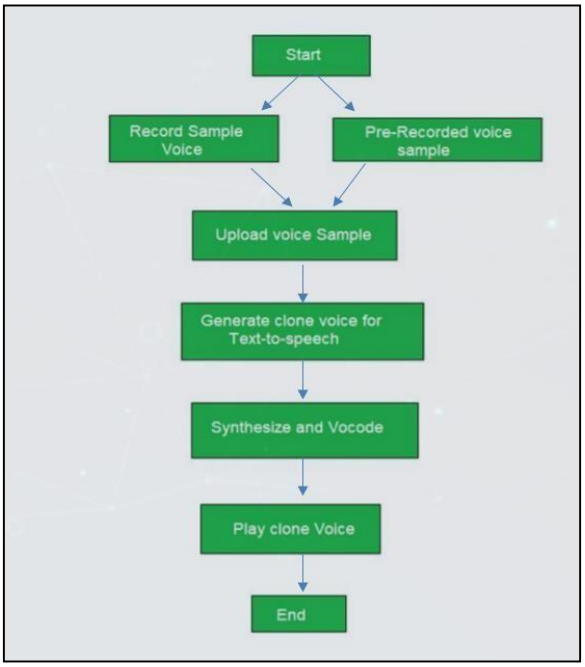
2.1 Introduction

It describes a new technique (often called Speech Vector to TTS, or SV2TTS) for taking a few seconds of a sample voice, and then generating completely new audio samples in that same style of voice. The SV2TTS model is composed of three parts, each trained individually. This allows each part to be trained on independent data, reducing the need to obtain high quality, multispeaker data. The individual components are: a) Speaker Encoder b) Synthesizer c) Wavenet vocoder.

2.2. Architecture/ Framework/Block diagram



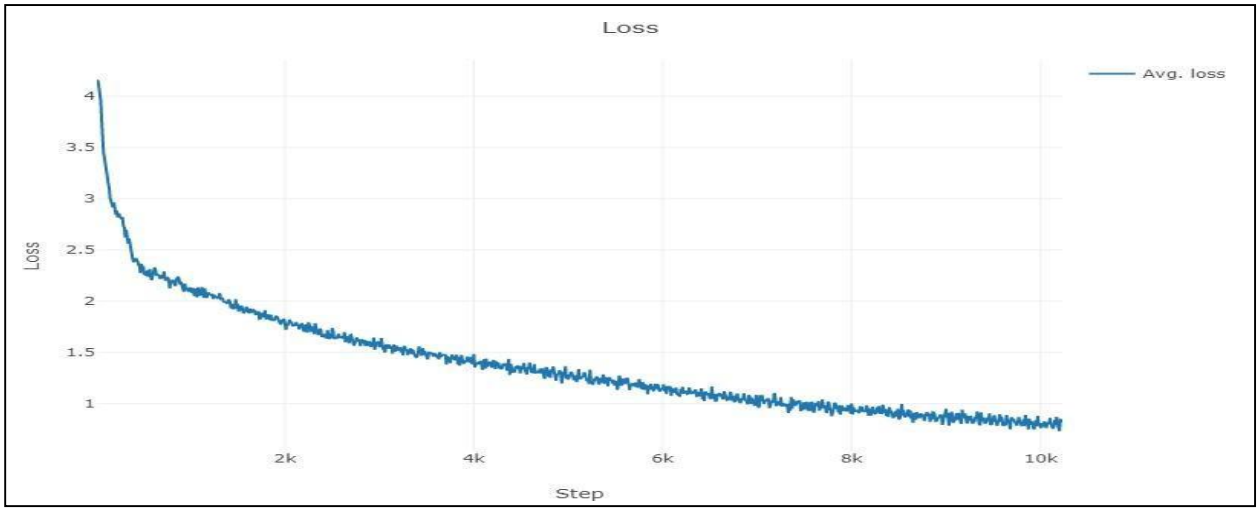
2.3. Algorithm and Process Design



2.4. Details of Hardware & Software

- Python 3.6 or 3.7 or 3.8
- Librosa == 0.8.1
- Matplotlib ==3.3.3
- Numpy ==1.19.3 Tensorflow ==1.5.0

2.5. Experiment and Results for Validation and Verification



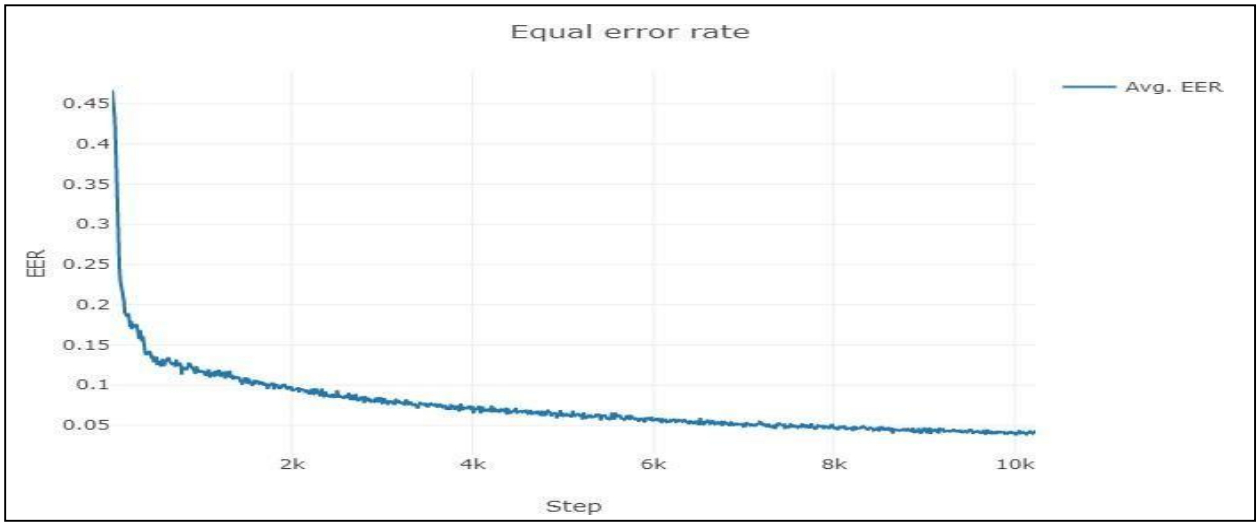


Fig 1. Loss and Equal Error Rate of 150 speakers while training till 10k iterations

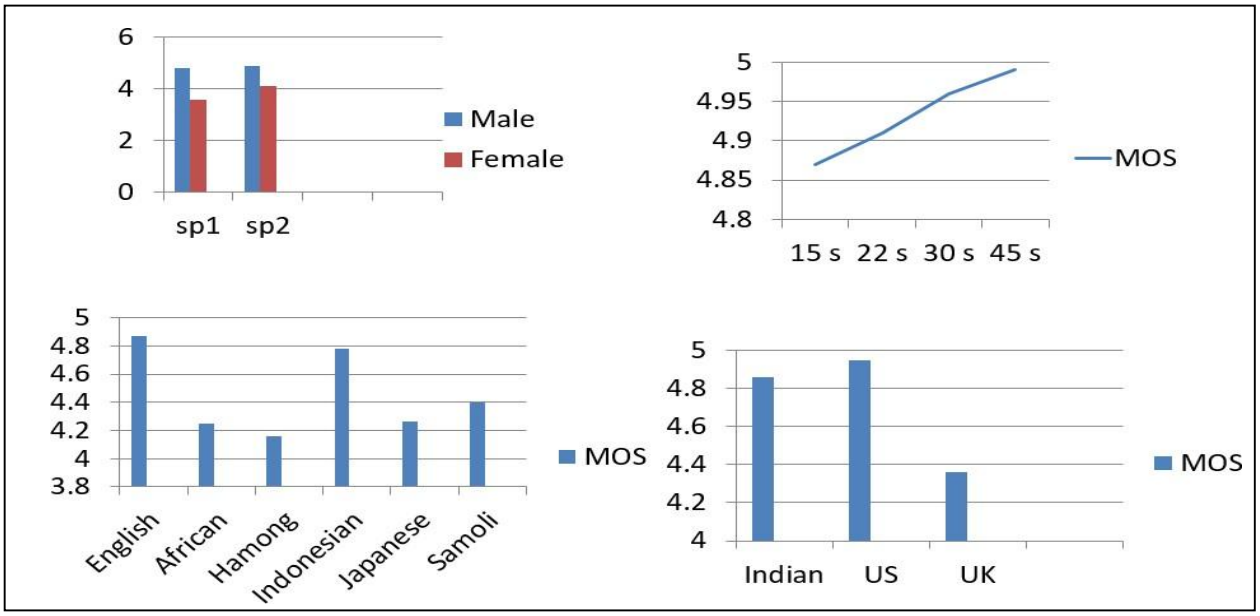


Fig 2. Mean Opinion Sore (MOS) Analysis of outputs

2.6 Conclusion

- Model performed better provided with more utterance per speaker.
- Accent of output of cloned voice is dependent on the accent present in dataset.
- Cloning of voice is independent of language.

References

[1] Ye Jia, Yu Zhang, Ron 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… Part of Advances in Neural Information Processing Systems 31 (NeurIPS 2018).

[2] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. . arXiv preprint arXiv:1609.03499, 2016.

[3] Fisher Yu, Vladlen Koltun, Multi-Scale Context Aggregation by Dilated Convolutions, Published as a conference paper at ICLR 2016

[4] Georg Heigold, Ignacio Moreno, Samy Bengio, Noam Shazeer, End-to-End Text- Dependent Speaker Verification, submitted to ICASSP 2016