INDEX

Index	1.
Chapter 1 – Introduction	2-3
1.1 Technology used	2
1.2 Field of Project	3
1.3 Technical Terms	3
Chapter 2 – Feasibility Study	4
2.1 Need and Significance	4
Chapter 3 – Methodology/ Planning of Work	5
3.1 System Design	5
Chapter 4 – Facilities Required for proposed work	6
4.1 Software Requirements	6
4.2 Hardware Requirements	6
Bibliography	7

CHAPTER 1 – INTRODUCTION

The goal of this work is to build a Text To Speech(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 an un-transcribed reference audio from a target speaker is used to synthesize new speech in that speaker's voice, without updating any model parameters.

Interestingly, speech naturalness is best rated with subjective metrics; and comparison with actual human speech leads to the conclusion that there might be such a thing as "speech more natural than human speech". In fact, some argue that the human naturalness threshold has already been crossed. Datasets of professionally recorded speech are a scarce resource.

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 modelling 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 multi-speaker 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 un-transcribed speech containing reverberation and background noise from a large number of speakers. We demonstrate that the speaker encoder and synthesis networks can be trained on unbalanced and disjoint sets of speakers and still generalize well. We train the synthesis network on 1.2K speakers and show that training the encoder on a much larger set of 18K speakers improves adaptation quality, and further enables synthesis of completely novel speakers by sampling from the embedding prior.

1.1 TECHNOLODY USED

Encoder-Decoder mechanism is used to have text to speech same to that of the natural voice given as an input. It is used to convert text into phonemes that are the values how the computer will read the text in particular input.

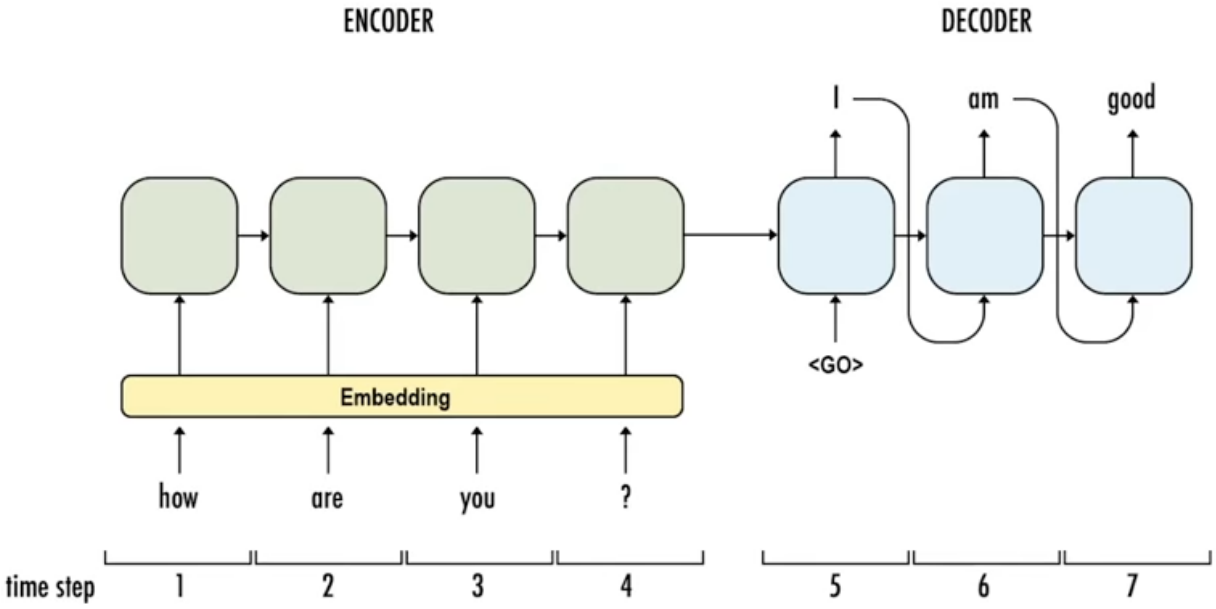

Figure 1: Encoder – Decoder System

Why? –"though" and "rough" should be pronounced so differently, even though they have the sane suffix. As such, we need to use a slightly different representation of words that reveal more information about the pronunciations.

White Room – [W, AY1, T, ., R, UW1, M, .] Crossroads – [K, R, AO1, S, R, OW2, D, Z, .]

1.2 FIELD OF PROJECT

ARTIFICIAL INTELLEGIENCE: Artificial intelligence (AI) refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions. The term may also be applied to any machine that exhibits traits associated with a human mind such as learning and problem-solving.

MACHINE LEARNING: Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

1.3 TECHNICAL TERMS

SPEECH NATURALNESS: As human being have natural sense of speaking, the machines are also required to produce that sense to make the voice generated from the machine look normal. It make the voice more natural to the listener and this will help to have the other person feel that the voice produced is the natural voice of the human only whose input was given.

SPEAKER SIMILARITY: There may exist a similarity in two people voice whose accent are same. The similarity of the voices will be a part of the project as the similarity in the speech can help the development. As the systems will be trained with the huge amount of data that will help the code to result into an efficient and more accurate voice generation. More the dataset of one particular accent, place, region etc people will be provided the better will the system created.

1.4 OBJECTIVES

- The goal of this work is to build a Text To Speech(TTS) system which can generate natural speech for a variety of speakers in a data efficient manner.
- Provide a secured and safe voice cloned speech.
- This system will help to build test to speech in a natural voice.
- Help the users to have the recording of their voice for number amount of data without actually recording the audio file.
- Will be a leading development for the other technologies to be build on Siri in users own voice, saving distinct languages through native speaker's input, or saving the voice of loved ones.
- Building robots with natural voice as studied in the scientific research module 2017.

CHAPTER 2 - FEASIBLITY STUDY

The objective of feasibility study is to determine whether proposed system is feasible. The feasibility is to determine in four aspects. These are:-

TECHNICAL FEASIBILITY: The project is based on with camera and it Detects Object using OpenCV library with use of certain kinds of Algorithms and notifies the other person on fall in Android Phone .Therefore it is very much favoured by the technology.

FINANCIAL FEASIBILITY: The project is based on camera based and few electronic components like ARDUINO which are affordable making it financially feasible to implement.

OPERATIONAL FEASIBILITY: This system notifies the person about falls happened to the detected person but lacks some operation by the user yet.

SCHEDULE FEASIBILITY: Schedule feasibility is measure of how reasonable project time table is the system is found schedule feasible because the system is designed in such a way that it will finished the prescribed time.

2.1 Need and Significance:

- 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.
- Needed for people who have accidently lost their voice. It will restore their ability to communicate naturally.

CHAPTER 3 - METHODOLOGY/ PLANNING OF WORK

3.1 SYSTEM DESIGN

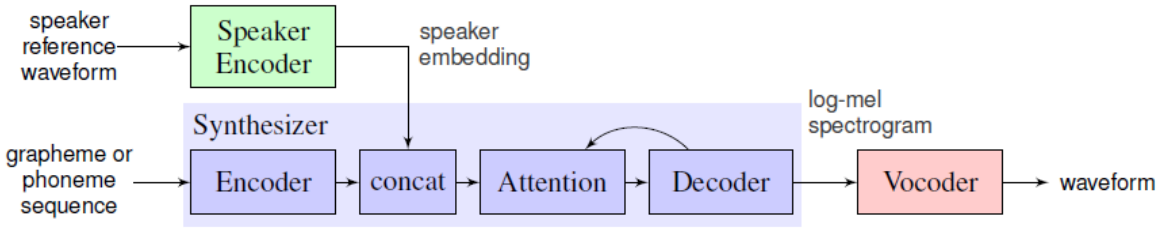

Figure 2: Multi-Speaker Speech Synthesis Model overview. Each of the 3 components are trained independently.

SPEAKER ENCODER: The speaker encoder is used to condition the synthesis network on a reference speech signal from the desired target speaker. Critical to good generalization is the use of a representation which captures the characteristics of different speakers, and the ability to identify these characteristics using only a short adaptation signal, independent of its phonetic content and background noise. These requirements are satisfied using a speaker-discriminative model trained on a text-independent speaker verification task.

SYNTHESIZER: An embedding vector for the target speaker is concatenated with the synthesizer encoder output at each time step. In contrast to, we find that simply passing embeddings to the attention layer, converges across different speakers.

NEURAL VOCODER: A vocoder to invert synthesized spectrograms emitted by the synthesis network into time-domain waveforms. The architecture is the same as that described in, composed of 30 dilated convolution layers. The network is not directly conditioned on the output of the speaker encoder.

CHAPTER 4 - FACILITIES REQUIRED

4.1 SOFTWARE REQUIREMENTS

- Operating system (Windows)
- python 3.7 or greater
- TensorFlow >= 1.1
- NumPy >= 1.11.1
- LibROSA == 0.5.1

4.2 HARDWARE REQUIREMENTS

- Computer Storage
- Ram>=4GB
- i5 Processor
- Microphone
- Speaker

BIBLOGRAPHY

https://thenextweb.com/artificial-intelligence/2018/02/26/baidus-ai-can-clone-your-voice-and-give-it-a-different-gender-or-accent/

https://www.youtube.com/watch?v=6KHSPiYlZ-U

https://www.youtube.com/watch?v=jPm05I5K05M

https://imgur.com/a/TRT0Z

https://medium.com/syncedreview/clone-a-voice-in-five-seconds-with-this-ai-toolbox-f3f116b11281

https://arxiv.org/pdf/1806.04558.pdf