THE ANDREW & ERNA VITERBI FACULTY OF ELECTRICAL ENGINEERING

Work in Progress

SIPL
Signal and Image Processing Lab

TECHNION Israel Institute of Technology

Speech-to-Singing Conversion Using Deep Learning

Omri Jurim and Ohad Mochly, Supervised by Yair Moshe and Gal Greshler

In collaboration with Lyrica

Introduction

- Few papers have been published in the science community regarding speech to singing (S2S) conversion, yet natural and accurate sounding output is hard to obtain
- In previous SIPL projects, speech to singing conversion had been partially achieved by using classic signal processing methods
- Converting speech to a song can help to memorize numbers, texts or can be done for entertainment purposes

Goals

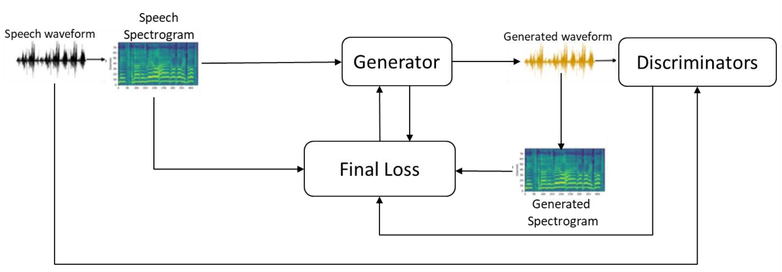
- Converting speech to singing using deep learning methods
  - Natural human sounding output
  - Full conversion process by deep learning methods
  - Time-domain-based conversion using generative adversarial network (GAN) architecture

Challenges

- No prior known solutions for S2S time-domain-based training
- Matching the data between speech to corresponding singing
- Deep network architecture which will satisfy the conversion (adapting existing network for singing+speech input and adding pitch to the training process)

HiFi GAN [Kong et al., 2020]

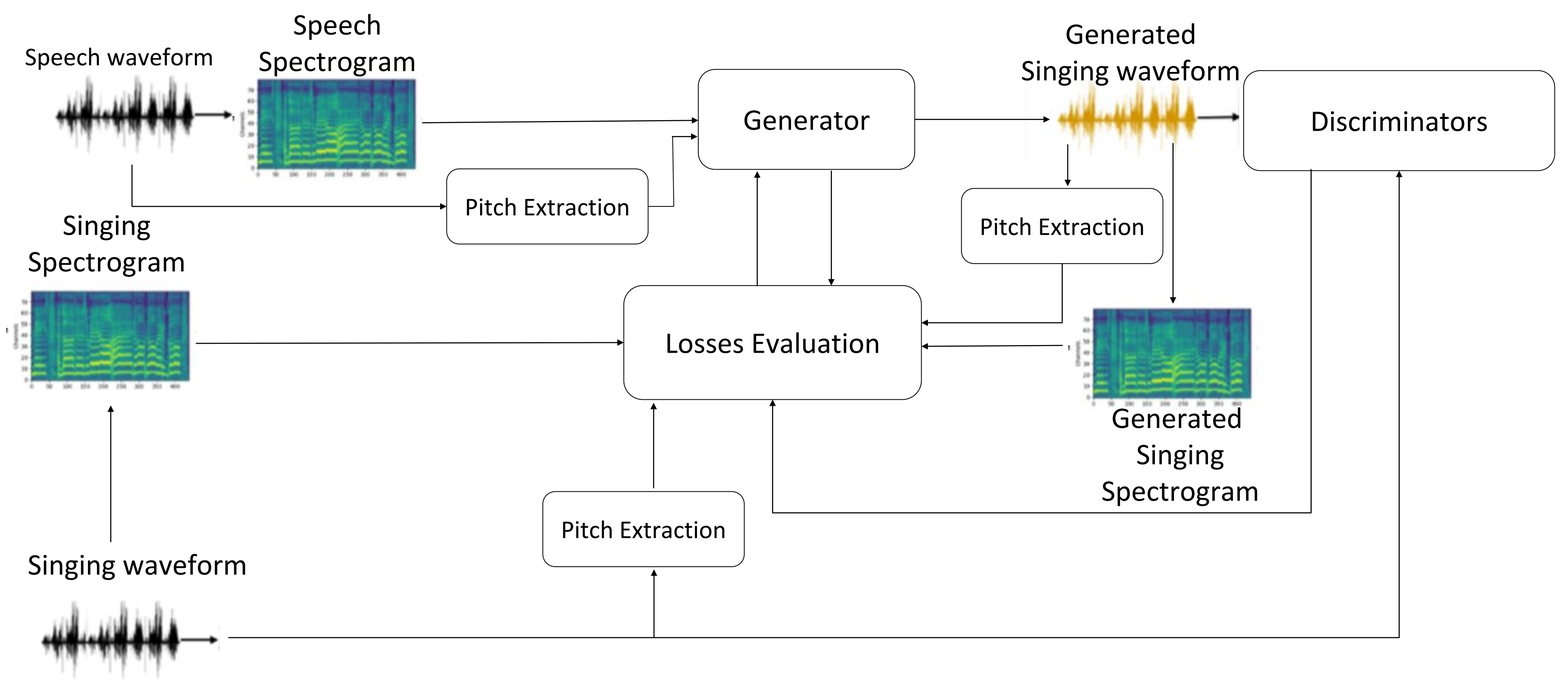
- Generates high fidelity waveform from spectrogram (speech synthesis)
- A state-of-the-art GAN architecture
- Evaluate the loss based on spectrogram and waveform (time domain)
- Achieves a higher MOS (mean opinion score) than the best publicly available models for speech synthesis



Core Training Solution Idea

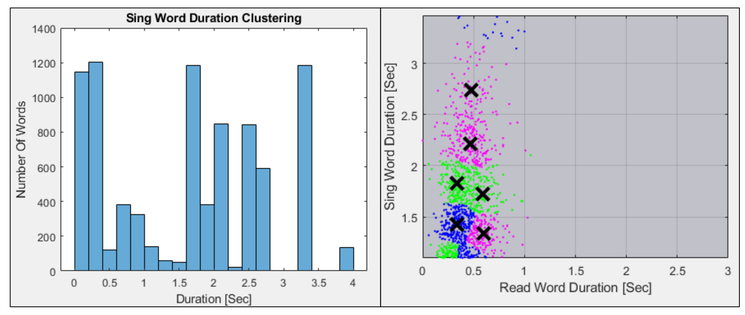
- The solution is based on HiFi GAN architecture
- Corresponding speech and singing words as inputs
- Loss is calculated using the waveform, spectrogram and pitch from generated singing and ground truth singing

Training scheme:



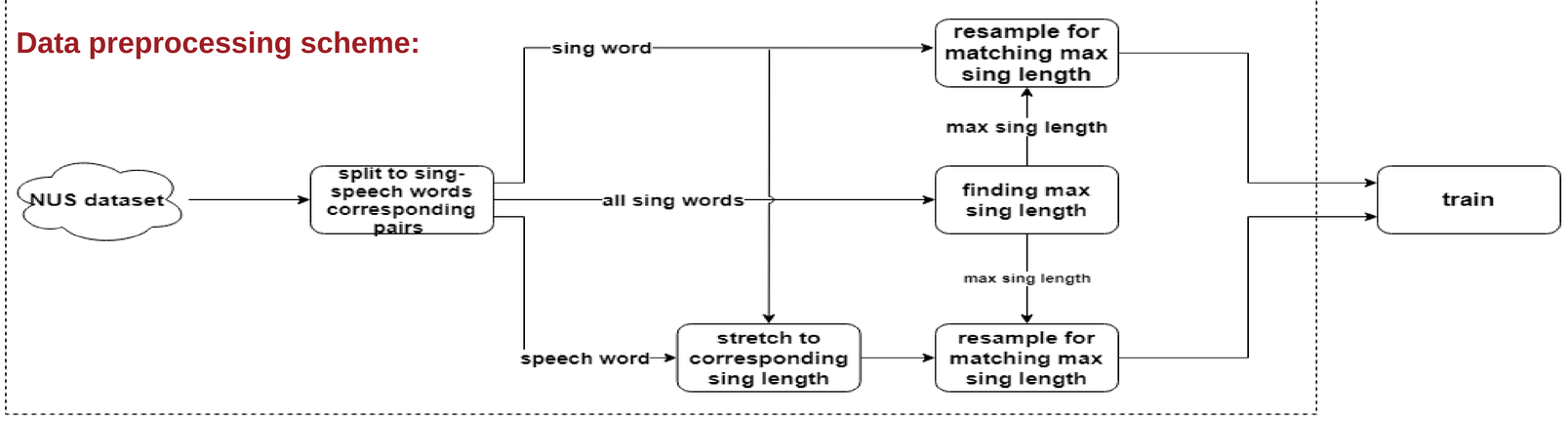
Data Preprocessing

- We used coupled speech & sing recordings from the National University of Singapore dataset
  - [Sharma et al., 2020]
  - ~50 songs and speech lyrics
  - ~170 minutes in total
  - Split songs and coresponding speech into single words
  - Split the data to subsets based on clustered song output durations



- We need a uniform length for all speech-singing pairs with minimal data loss/distortion:
  - Time stretch the shorter of the pair to the longer's size
  - Resample both to the size of the longest file in the dataset

Data preprocessing scheme:



Results So Far

- Successfully trained the HiFi GAN model on singing inputs – transfer learning on the pretrained model

June 2021