End-to-end Text to Speech Synthesis

CS 229 Final Project Autumn 2018

Xiao Wang (xiao1105), Yahan Yang (yangy96), Ye Li (liye5)

Introduction

- ***Motivation:*** One of the most challenging problem in audio/music processing is text-to-speech synthesis. With rapid development of deep learning, researchers invent many end-to-end algorithms for real life problems, which leads more innovative methods in solving speech synthesis problem.
- ***Problem definition:*** Generate audio file from text input.
- ***Approach:*** Combine seq2seq model with attention mechanism. Also try simple models as baseline.
- ***Challenge:*** This problem require complicated and well-designed model to generate audio reasonable files.

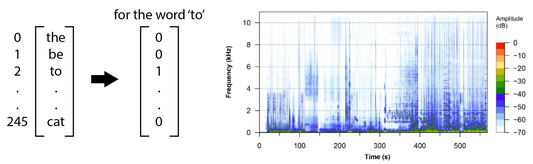
Preprocess of dataset

Dataset

- ***LJ Speech Dataset***. This is a public domain speech dataset consisting of 13,100 short audio clips of a single speaker reading passages from 7 non-fiction books.

Text:

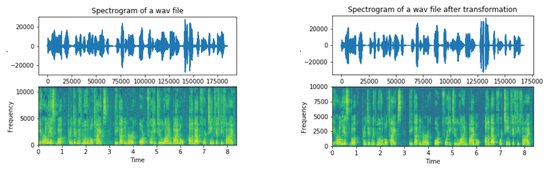
- ***CMU Dictionary.*** Every word has a potential of being transformed into phonemes. Build a symbols dictionary with phonemes and alphabet.



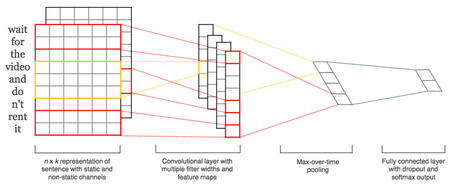
Audio

- ***FIR filter.*** A finite impulse response (FIR) filter is a filter structure that can be used to implement almost any sort of frequency response digitally. It can smooth noise.

- ***Short-time Fourier transform***. STFT, is a Fourier-related transform used to determine the sinusoidal frequency and phase content of local sections of a signal as it changes over time

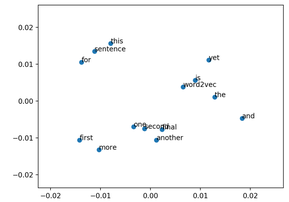


Models



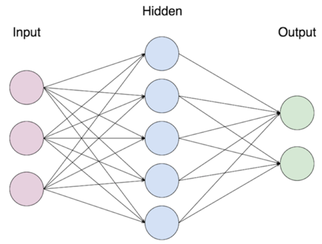
Word Embedding Model

words or phrases from the vocabulary are mapped to vectors of real numbers. We prepend this embedding layer to each of our models.



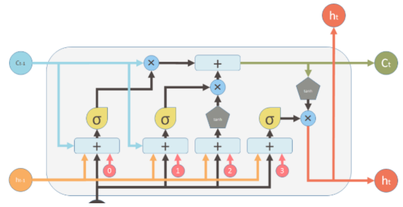
SVR model

The program generates a SVR for each timestep, so the total number of SVR in our model equals to the number of time step after we preprocess data. We tried a linear kernel and a polynomial kernel for our SVM models. In each time step, we have a support vector multi-regressor for training each input text matrix and an array of spectrum output.



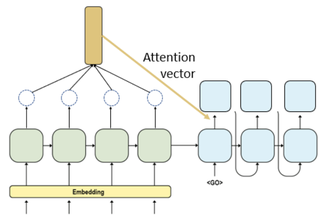
Neural Network model

To linearize the output, the 2D array was reshaped to a 1D array first, and treat the long 1D array as the output predictions for this model. The input layer is simply the entire vector of sequences, and the hidden layer has the same number of neurons as input length, which is fully connected to the inputs.



Seq-2-seq model with attention

Input data was first put into embedding layer to align the dimension with output wave matrix. And then output of embedding layer was feed into bidirectional LSTM encoder model. The hidden state and cell state of attention bidirectional LSTM model were shared with decoder model.



Postprocess of dataset

- After the system generates output spectrum matrix from prediction, we utilize inverse FIR filter to transfer spectrum to audio signal and save the wave into byte with the method provided in spicy library.

Results

- We put our sample wav files in the web (https://www.xiaowang.me/cs229).
- Evaluation of our machine learning algorithm is depending on the naturalness, which is given by a group of native speakers at Stanford University and calculate the mean opinion score(MOS) as a standard.
- After asking 10 Stanford students, we obtained the mean score of (1.0, 1.7, 2.5) for these three models, respectively.

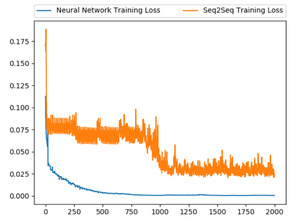
Label	Excellent	Good	Fair	Poor	Bad
Rating	5	4	3	2	1

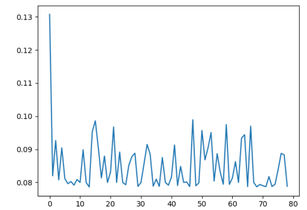
	SVR	NN	Seq-2-seq
MOS	**1**	**1.7**	**2.5**

Performance & Analysis

SVR model

The wave file generated from our SVR model mainly consisted of disjoint words, so that it does not sound like consistent human speech. One problem of SVR model is that the training time is too long, since the model works with many separate models and a number of multi-regressors.





Simple Neural Network

The wave file generated from this simple neural network does not sounds like consistent speech. It was just a random combination of phonemes and words. Although through the training process, the loss function can be minimized down to relatively low magnitude, when comes to the dev/test data, the result sounds not quite reasonable.

Seq-2-seq model with attention

The wave file generated from this model sounds like human speech. Even though after training, the loss function can be minimized down to the magnitude of 10e-3, when comes to the dev/test data, the result sounds not quite reasonable.

Reference

[1] Wang, Yuxuan, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang et al. "Tacotron: Towards end-to-end speech synthesis." *arXiv preprint arXiv:1703.10135* (2017).

[2] Perraudin, Nathanaël, Peter Balazs, and Peter L. Søndergaard. "A fast Griffin-Lim algorithm." In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*, pp. 1-4. IEEE, 2013.

[3] Sotelo, Jose, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio. "Char2wav: End-to-end speech synthesis." (2017).

[4] Arik, Sercan, Gregory Diamos, Andrew Gibiansky, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. "Deep voice 2: M