Using Tacotron to Emulate Voices

Eidan Maimoni
Department of Computer Science
New York University, NY, USA
emm792@nyu.edu

Sakawkarn Piyawitwanich
Department of Computer Science
New York University, NY, USA
sp4887@nyu.edu

***Abstract*—**

This paper is an application of Tacotron, a neural network model for text-to-speech synthesis, on a specific dataset. We theorize that we can emulate and recreate any speakers voice to a degree of accuracy by taking advantage of how the model learns from audio samples along with their small nuances in pronunciation. In our case, rather than training over a large data set of various voices of different genders and accents we instead apply only recurring audio samples of British women. This instead generates a more specific character voice rather than a much more generic english. We show that there is evidence that an accurate replica of the speakers voice is possible.

Keywords—analytics, machine learning, deep learning, Tacotron, text-to-speech

I. INTRODUCTION

Applications of machine learning are reaching various areas rapidly. With the use of neural networks it seems as if almost anything is possible. Our project paper researched recent publishings of advanced methods that learn and predict sound well. Specifically, that of human text-to-speech. Inspired by the original WaveNet paper from class we looked further into the work that has been done by Google and came across Tacotron and Tacotron 2. While WaveNet is a deep neural network for generating raw audio waveforms [2] it still requires expertise in the field of speech and language. Requiring inputs such as sequences of linguistic and phonetic features. Alternatively, Tacotron does not require such features as input. It generates speech simply from text-audio pairs [1]. Finally, we come across the state of the art Tacotron 2. A neural network architecture that combines the best of both underlying models of WaveNet and Tacotron for both their impressive results and simplicity respectively [3].

Our goal was to apply Tacotron 2 with a specific dataset to capture the embodiment of the speaker rather than just the ability to speak. Initially we had hopes of replicating cartoon characters voices due to the availability of audio clippings and subtitles. However, this resulted to be too difficult due to licensing and extra work needed to prepare the data. Next, we moved on to the idea of audiobooks and their speakers but we encountered the same issues. More specifically, the lack of documentation of timing of the audiobook to the sentence level recorded. Instead, we decided on a simpler approach which is that of a "generic" British woman by utilizing the large prepared data sets of VCTK-Corpus and M-AILABS. By implementing audio-text pairs of British women book reading recordings we hoped to generate an output that could capture the nuances of the British accent along with that of the femininity of the voices.

Finally, due to issues with being able to train the WaveNet model by lack of proper equipment on our part. We regressed back to only the application of Tacotron alone for our project. With our output we observe how Tacotron managed to capture the elements we had hoped for from the data so quickly in the span of only 50 steps. Later, with the training of 100 steps it begins to develop further into more comprehensible speech as expected.

Coincidentally, we come across and see from other people's work that our idea carries. In the work of Kyubyong's Github repository for Tacotron [4] we can see how his training with a Nick Offerman audiobook mimics the actor's voice with a high level of clarity and similarity. Moreover, as of recent a team at AI company Dessa has developed a model known as RealTalk that recreates famous podcast host Joe Rogan's voice in a similar manner [5]. This project captures the essence of our goal that was to show how the goal of mimicking someone is indeed possible.

II. MOTIVATION

This application is of very high interest as it is often overlooked the possibilities of use of generating audio with machine learning. While WaveNet shows that a clearer and more authentic voice can be created, it does not delve into the area of emulating specific voices. A large use for this technology would be in the case of voice acting. Where in many cases a voice actor has to abandon a show leads to production abandoning the character, we can instead see the possibility of simply using the generative model to replace the voice actor completely to avoid this problem. On the other hand, it is also important to understand how well this technology works with there being a risk of it being used to falsify audio recordings. Attempting to have people say things they have never actually said. By understanding it we can create solutions to stopping such cases.

III. Related Work

The first paper we read was that of WaveNet [2]. This was a breakthrough paper as it applied dilated casual convolutions in its model to train the neural network. This first would generate human language like sounds without coherence. While picking up on realistic speaking characteristics such as breathing. Next they conditioned the speech on text with linguistic features. Thus, creating the high scoring 5-scale MOS level of naturalness.

The next paper we researched was in attempt to bypass the need for these linguistic features as input. Here we are introduced to Tacotron [1]. Utilizing a seq2seq model, encoders, decoders, recurrent neural networks, the architecture takes the input text and audio and produces spectrogram frames. These spectrogram frames are then converted into the audio wave forms we desire through a Griffin-Lim algorithm with Tensorflow. Once again achieving a high scoring MOS.

The last paper we researched was the most advanced. Tacotron 2 synthesizes speech in a similar approach to Tacotron requiring only text and audio, however better [3]. While Tacotron achieves a MOS score of 3.82 and WaveNet of 4.21, Tacotron 2 reaches as high as 4.53. This system unifies both aspects of Tacotron and WaveNet. Generating mel spectrogram with the seq2seq which is then passed to the WaveNet vocoder for waveform output. Applying this model however was not as simple.

IV. Application Design

Here we see the Tacotron and Tacotron-2 diagrams respectively. We can see the main difference where Tacotron uses a Linear-scale spectrogram whereas the Tacotron-2 relies on the Mel Spectrogram to be passed into the WaveNet model.

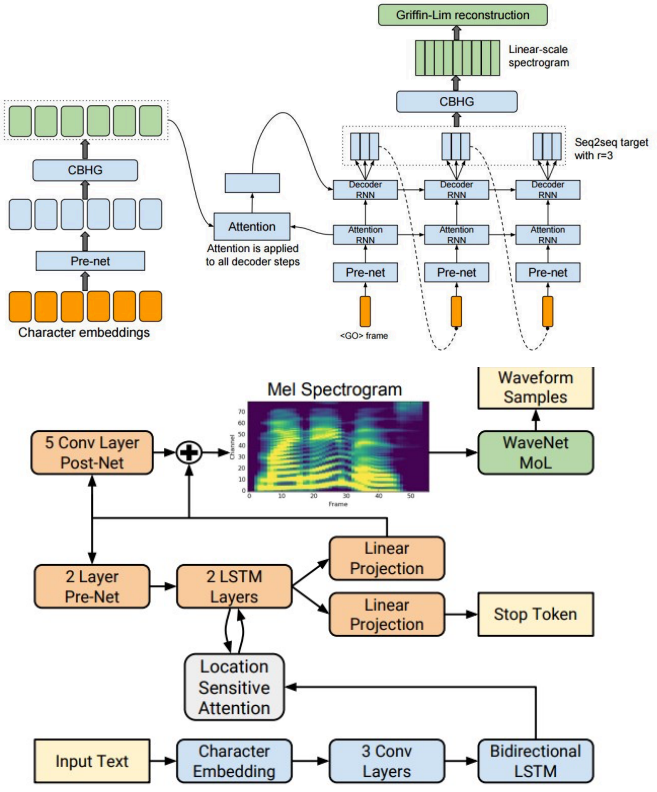


Fig. 1. Block diagram of the Tacotron 2 system architecture.

V. Datasets

There are two datasets used. One from CSTR VCTK-Corpus and the second is the M-AILABS Speech Dataset.

The first one, the VCTK-Corpus dataset, is provided by the University of Edinburgh. This dataset contains audio files of 109 native English speakers of which read around 400 sentences each, as well as the corresponding transcripts. The total size of the dataset is 15.57 GB.

The second one is the M-AILABS dataset which also contains the audio files of varying length from 1-20 seconds and their transcripts. The M-AILABS data is based on LibriVox and Project Gutenberg. The audio in this dataset was recorded by LibriVox project and was published to the public in 1884 to 1964. The data from M-AILABS that we used in this project are the female voices in U.K. English. The size of the dataset is 4.9 GB.

VI. Experiments

Data preparation

For the VCTK-Corpus dataset, we have downloaded the whole dataset containing the recordings of English native speakers of various regions of the world and their text. We used only the data from people who are labeled to have a British English accent. Moreover, each speaker has multiple recordings of audio files (.wav) and corresponding text files (.txt). We combined all the text files together to create a metadata in CSV format that includes the corresponding audio files label name. Lastly, we moved all the audio files to one directory called "wavs".

For the M-AILABS dataset, we downloaded only the part of the dataset that was of Female British English speakers. The dataset is already formatted for us to use accordingly.

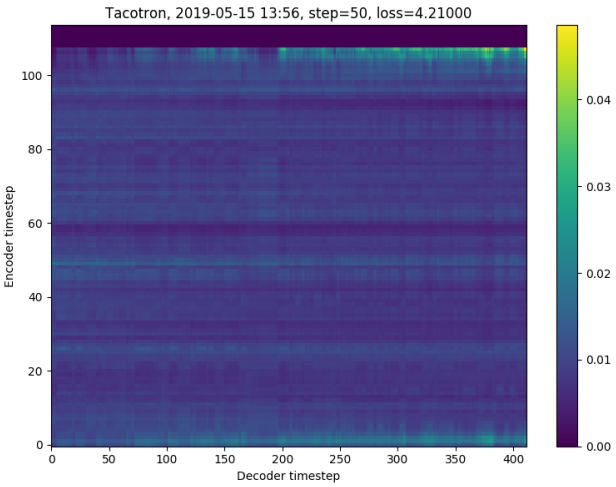
With the datasets prepared and finished we now proceed to the preprocess of the data.

Preprocessing

We run preprocessing on each dataset separately to prepare NumPy array files (.npy) that can be used in training the model. The data after the preprocessing are stored in training data directory.

Model training

We originally intended to train the Tacotron 2 model which trains the datasets on both Tacotron, to produce mel-spectrogram, and WaveNet, to turn mel-spectrogram into waveform sound. We tried to train the WaveNet model initially for 50 and 100 steps. Unfortunately, we could not supply the GPU hardware required to train WaveNet model as it required a Nvidia GPU for CUDA as well as Tensorflow-gpu. Instead pivoted and only trained and used the Tacotron model. The model alignments of Tacotron model at 50 and 100 steps are showed in Fig. 2-3 below. Similarly, their mel-spectrograms of target and prediction for 50 and 100 steps are also showed in Fig. 4-5 below.



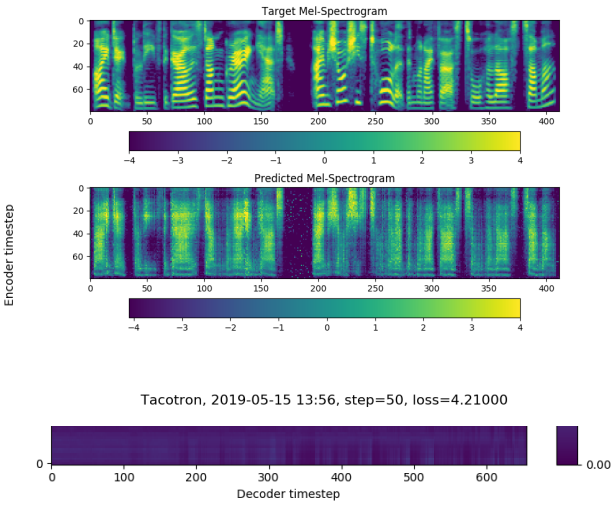


Fig. 2. Tacotron model alignment at 50 steps of training.

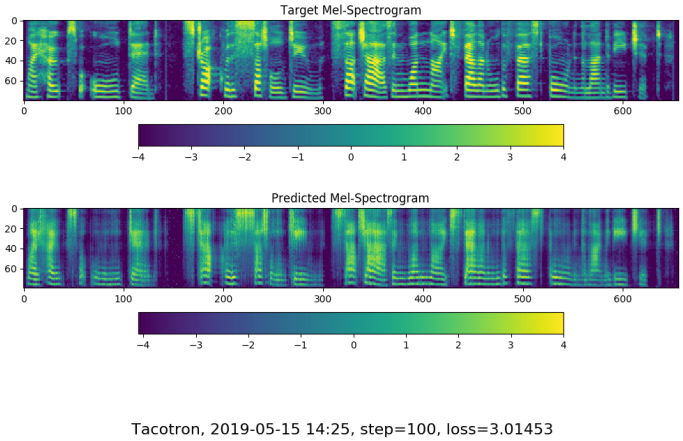
Fig. 3. Tacotron model alignment at 100 steps of training.

Fig. 4. The mel-spectrogram of target and prediction of Tacotron model at 50 steps of training

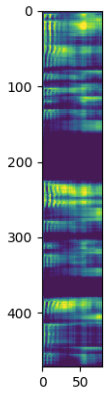
Fig. 5. The mel-spectrogram of target and prediction of Tacotron model at 100 steps of training

After training the Tacotron, we find the outputted completed model in the "logs-Tacotron" directory. The 100 step waveform audio clips were created as well and can be listened to in the wavs directory. The sound that has been produced by the Tacotron model is not completely clear due to the non-uniformed noises. But we can recognize the English accent as well as the female voice if listening with close attention.

Audio Synthesizing



To predict the spectrogram of the sound, we ran the audio synthesis on the mel-spectrograms that we derive from the Tacotron model. The synthesis, known as the Ground Truth Aligned synthesis, provides the mel-spectrogram prediction that can be further used in training the WaveNet vocoder as defined in the Tacotron 2 paper. The results of running the audio synthesis are the NumPy array file (.npy) of the predicted



individual mel-spectrograms from each audio file of the speaker. The prediction of mel-spectrogram from Elizabeth Klett, one of the speakers, reading a sentence from a book named Jane Eyre is showed in Fig. 6 below.

Fig. 6. The example of prediction of mel-spectrogram of jane_eyre_38_f000090 after the synthesis

We could then use the mel-spectrograms to produce the waveform sound if we train WaveNet model. But as we have mentioned before that due to the lack of hardware requirements and preparation, we could only get the predictions of mel-spectrograms. We hope to build a complete Tacotron 2 application with more meaningful results if given the chance again.

VII. Conclusion

We used audios and transcripts of Female English speakers with British accent and implemented a Tacotron model to predict the mel-spectrograms frames that can be used as an input to the WaveNet model to output waveform sounds. The Tacotron model we have trained used only 100 step sizes, combined with audio synthesis it took approximately 30 hours to complete on a CPU. We encounter results that confirm our theory albeit weakly. But with evidence there is motivation to try again for better.

VIII. Future Work

Unfortunately, training such models and gathering data is not easy. In the future we hope to be able to take audio recordings from a show or movie, along with it is following subtitles. By utilizing the subtitles timings and cutting the data into smaller audio lengths accordingly we hope to have just as large and convenient of a dataset as our project. With the more interesting and larger dataset we can take advantage of higher performing computers and time windows to properly train higher quality trained models with more convincing outputs. Ideally we train the model with at least 200,000 steps for better results in terms of clarity. If given the appropriate hardware, we could have also trained WaveNet model to complete the Tacotron 2 application for maximum results. However, we hope that our implementation can be further used in the future project to produce the consistent sound that we initially intended to do.

Acknowledgment

We would like to thank LibriVox and Project Gutenberg for the M-AILABS dataset, as well as the University of Edinburgh for the VCTK-Corpus dataset. Further thanks to Rayhane Mama for his Tacotron 2 Github repository. Finally, also we would like to thank Professor Rajesh Ranganath for his teachings and original inspiration for this project with his readings.

References

1. Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark: "Tacotron: Towards End-to-End Speech Synthesis", 2017; arXiv:1703.10135.
2. Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior: "WaveNet: A Generative Model for Raw Audio", 2016; arXiv:1609.03499.
3. Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis: "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions", 2017; arXiv:1712.05884.
4. Kyubyong, Tacotron, (2018), GitHub repository, https://github.com/Kyubyong/tacotron.
5. Dessa. "RealTalk: This Speech Synthesis Model Our Engineers Built Recreates a Human Voice Perfectly (Part..." *Medium*, Medium, 15 May 2019, medium.com/@dessa_/real-talk-speech-synthesis-5dd0897eef7f.