Title: Interim Report

G005 (s1862323, s1826377)

Abstract

The project seeks to improve the quality of Corentin Jemine's open source implementation of SV2TTS, which will be evaluated using subjective and objective metrics. Baseline experiments have been conducted to establish a baseline performance, with the primary goal of improving the quality of cloned voices of TTS, and the secondary goal of reducing the complexity and time for generating these voices.

1. Introduction

Deep learning has the potential to revolutionise the field of Text-To-Speech (TTS) with its powerful ability to learn the complex relationship between text and corresponding speech. Besides, deep learning methods have a wealth of possibilities if voice cloning. Combining TTS and voice cloning, one's voice of speaking of text contents can be generated without its presence.

Text-to-Speech (TTS) is a rapidly growing field of research and development, with numerous practical applications in domains such as accessibility, entertainment, and education. The aim of TTS is to produce natural-sounding speech from written text. To achieve this goal, researchers have been using a variety of techniques, such as natural language processing and speech synthesis, to generate more accurate and lifelike audio. This technology has been used in a variety of applications such as virtual personal assistants and digital storytelling. This technology has been used in a variety of settings, ranging from voice-based assistants to automated customer service. As technology continues to improve, it is expected to be increasingly used in other areas, such as healthcare and education, to provide more efficient and accessible services.

Voice is an essential part of humanity, a fundamental means of communication, and a vital link between people. Unfortunately, there are many children who are not accompanied by their parents and whose voices are rarely heard. .(Wiki, 2022) These children can be found in impoverished areas of China(Economist, 2021), India(Express, 2018), the Philippines, and Pakistan, among other countries. One of the reasons for this is the lack of stable employment opportunities in these locations, forcing their parents to seek work in other places to earn enough money to provide for their families. The lack of parental guidance and presence can have a detrimental effect on these children, limiting their educational opportunities and access to basic needs. This can lead to a cycle of poverty and deprivation, making it difficult for them to break out of the situation they find themselves in.

Our project aims to develop a Text-To-Speech (TTS) system that can precisely replicate the user's desired vocal style, tone, and so on. This TTS system can provide high-quality vocal companionship for children and even enable people to recreate the voices of their lost loved ones, thereby allowing them to experience their presence and companionship, despite their physical absence. This is a highly meaningful and powerful technology that can bring a great deal of comfort to those who require it. It is our hope that this technology can be used to make a positive difference in the lives of people from all walks of life. We believe that this system can be used to improve the quality of life for people in need, such as those who are socially isolated, disabled, and/or have lost a loved one. We are committed to ensuring that this technology will be accessible and affordable to all those who could benefit from it.

Despite the advances made in the field of text-to-speech (TTS) technology, there remain a number of challenges that need to be addressed. This includes the need to acquire high-quality, diverse data sets in order to create more advanced and flexible TTS synthesis models. There is also a necessity to incorporate additional linguistic and paralinguistic information into current TTS systems to further improve the quality and accuracy of the generated output. It is hoped that with these enhancements, further development in the TTS field can be achieved and more realistic and convincing speech synthesis can be produced.

Some related works in TTS include Tacotron: Towards End-to-End Speech Synthesis (Wang et al.), which demonstrated the feasibility of generating high-fidelity speech from text alone; Deep Voice: Real-time Neural Text-to-Speech(Arik et al., 2017), which showed that a single neural network can learn to generate speech from a large corpus of data; Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis, which demonstrated the effectiveness of transfer learning for TTS applications.(**?**) These works have set the stage for further investigation into the field of TTS and have enabled the development of more advanced architectures and methods for generating realistic and natural-sounding speech from text.

SV2TTS introduced a structure of models for TTS with voice cloning, which is used as the baseline to improve.

The project aims to improve the open source implementation by Corentin Jemine (Jemine, 2019) of SV2TTS (Jia et al., 2018) to the next-level quality of generated voice. The quality will be evaluated by subjective and objective metrics. A set of baseline experiments have been performed to get the baseline performance. The core objective is to improve the quality of cloned voice of TTS evidently. The secondary objective is to reduce the complexity and time for generating cloned voices of TTS.

2. Data set and task

The system consists of three independently trained components: The Generalized End-To-End Loss for Speaker Verification (GE2E) Encoder, Tacotron Towards End-to-End Speech Synthesis synthesizer, and WaveRNN Efficient Neural Audio Synthesis vocoder. Three models all plan to use the LibriSpeech ASR corpus-100h dataset for model training (Korvas et al., 2014). This dataset, compiled by Vassil Panayotov with the support of Daniel Povey, consists of approximately 1000 hours of 16kHz read English speech. The data was sourced from the audiobooks available in the LibriVox project and has been carefully cut, segmented and aligned.(Korvas et al., 2014) Moreover, the LibriSpeech dataset has been extensively tested for accuracy, and has been found to be suitable for various speech-related tasks such as speech recognition, speech synthesis and speech-based dialogue systems. This makes it an ideal choice for training deep learning models.

The baseline of our system is based on a 100-hour clean dataset which serves as the foundation for our test set. By combining the clean set with other test sets, we are able to thoroughly evaluate the performance of our system. In the near future, we plan to expand the size of our training set, as well as classify it by age group to identify different styles. Additionally, we are looking to make improvements to our style acquisition for the dataset of parents in order to identify the style of different age groups accurately. Furthermore, we are also aiming to enhance the imitation learning ability of our system for style so that it can produce even more accurate results. To this end, we are exploring ways to improve the learning algorithms and the quality of our datasets, in order to increase the accuracy and improve the overall performance of our system.

3. Methodology

Corentin Jemine reproduced the framework introduced by SV2TTS as an open-source public implementation. Jemine's public implementation is set as the baseline of this project. The baseline experiment is going to determine the quality of generated results from the perspective of research problem, using TTS with voice cloning imitate human speaking text messages.

The SV2TTS structure consists of three independently trained components: (1) a speaker encoder network, which is trained on a speaker verification task using an independent dataset of noisy speech without transcripts from thousands of speakers. This network is able to generate a fixed-dimensional embedding vector from only seconds of reference speech from a target speaker. This is then used to condition (2) a sequence-to-sequence synthesis network based on Tacotron 2, which is capable of generating a MEL spectrogram from text, conditioned on the speaker embedding. Finally, (3) an auto-regressive WaveNet-based vocoder network is used to convert the MEL spectrogram into time domain waveform samples. The entire system is a powerful tool for text-to-speech synthesis, allowing for the generation of natural-sounding speech from text (Jia et al., 2018).

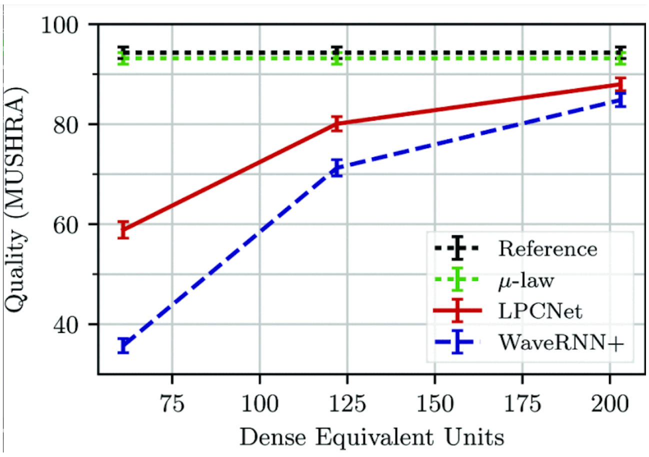

Figure 1. Performance of LPCNet

We decide to improve the third component, WaveNet-based vocoder, by replacing WaveRNN vocoder with LCPNet vocoder (Valin & Skoglund, 2018). For the vocoder, the implementation by Corentin Jemine adopts the "alternative WaveRNN"(Jemine, 2019) model, an improved WaveRNN model (Kalchbrenner et al., 2018). LCPNet have shown better performed quality comparing with WaveRNN under the same complexity (Kalchbrenner et al., 2018).

4. Experiments

We generated 27 samples from 14 speakers via Jemine's implementation to compare with the ground truth. The data are from LibriSpeech's test set (LibriSpeech-test-clean), each sample is generated with a voice embedding sample and text content of target sample. The target sample is the ground truth to compare with, and the embedding sample is another piece of voice from the same speaker. The objective evaluation metric we use is Mel-Cepstral Distortion (MCD) with Dynamic Time Warping (DTW). The average MCD of the 27 pairs of samples (generated, groun truth) is 7.624 (penalty:0.338, 716.222) For further experiments, we will mainly use MCD to determine our improvements and use Mean Opinion Score (MOS) as secondary subjective evaluation metrics.

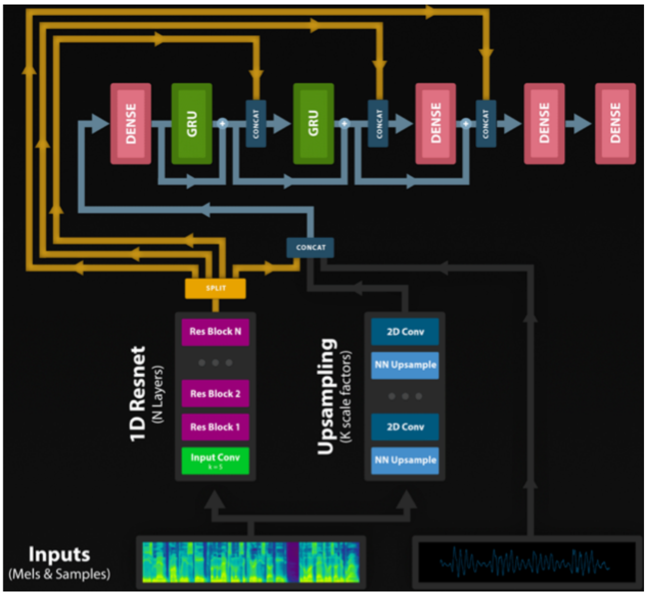

Figure 2. Structure of WaveRNN used in Jemine's public implementation

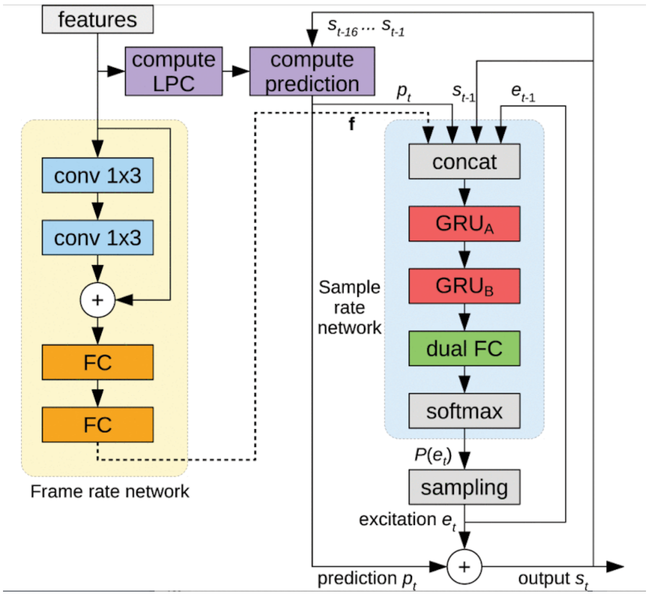

Figure 3. Structure of LPCNet

5. Interim conclusions

After using the test-clean database sound and my own sound as inputs, I conducted a comparison between the two sounds. I found that the sound I generated was more prone to background noise, which had a noticeable effect on the Mel spectrum. This was in contrast to the reading of the database sound, which was not as affected by noise. Consequently, I believe that pure recording without any kind of noise processing could be the cause of the noise points I observed in my generated sound, and I am determined to improve it in the future. In order to do so, I plan to analyze the sounds more thoroughly and make adjustments accordingly. By doing this, I hope to be able to reduce the amount of noise present in the generated sound and make it indistinguishable from the database sound.

When the input audio is longer, it has a negative effect on the result of style imitation, and the generated sound is more like a machine sound; Similarly, when the text and generated sound are longer, the sound is more like a machine sound.

Also, the generated voice is often noised by current-like sound, which may be affected by vocoder(Mel spectrum to waveform).

6. Plan

1. Future plan: - Implement to replace WaveRNN vocoder with LPCNet vocoder - Evaluation by objective (MCD) and subjective (MOS) metrics.

2. Potential risk. -The model is hard to improve. The quality of the generated voice cannot be improved. -The implementation of LPCNet vocoder failed to initegrate with Jemine's open-source implementation. -The pre-trained models are not compatible with the new LPCNet vocoder. Re-training all models may requires weeks (not enough time and resources). -No enough resourses and time to perform subjective evaluation(MOS)

3. Backup plan. We can improve the quality of the generated voice to make it more similar to the desired style. This includes reducing background noise like wind, storms, and currents; adjusting the generated voice speed to match the talker's speed, and making it sound more like the real speaker. In addition to this, we want to bring the system closer to our motivation: to allow the user to hear the sound they want to hear and to be accompanied by a high-quality sound.

Additionally, we can instead improve the time or computing efficiency while keeping the level of quality.

References

Arik, Sercan O., Chrzanowski, Mike, Coates, Adam, Diamos, Gregory, Gibiansky, Andrew, Kang, Yongguo, Li, Xian, Miller, John, Ng, Andrew, Raiman, Jonathan, Sengupta, Shubho, and Shoeybi, Mohammad. Deep voice: Real-time neural text-to-speech, 2017. URL https://arxiv.org/abs/1702.07825.

Economist. The plight of china's "left-behind" children. 2021. URL https://www.economist.com/china/2021/04/08/the-plight-of-chinas-left-behind-children.

Express, The Indian. The children left behind, unesco report highlights the gaps in education policy for children of migrants. 2018. URL https://indianexpress.com/article/opinion/columns/unesco-migration-india-the-children-left-behind-editorial-5471538/.

Jemine, Corentin. Master thesis: Real-Time Voice Cloning. 2019. URL https://matheo.uliege.be/handle/2268.2/6801. Unpublished master's thesis.

Jia, Ye, Zhang, Yu, Weiss, Ron J., Wang, Quan, Shen, Jonathan, Ren, Fei, Chen, Zhifeng, Nguyen, Patrick, Pang, Ruoming, Moreno, Ignacio Lopez, and Wu, Yonghui. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. 2018. doi: 10.48550/ARXIV.1806.04558. URL https://arxiv.org/abs/1806.04558.

Kalchbrenner, Nal, Elsen, Erich, Simonyan, Karen, Noury, Seb, Casagrande, Norman, Lockhart, Edward, Stimberg, Florian, van den Oord, Aaron, Dieleman, Sander, and Kavukcuoglu, Koray. Efficient neural audio synthesis. *2018*, 2018.

Korvas, Matěj, Plátek, Ondřej, Dušek, Ondřej, Žilka, Lukáš, and Jurčíček, Filip. Free English and Czech telephone speech corpus shared under the CC-BY-SA 3.0 license. In *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC 2014)*, pp. To Appear, 2014.

Valin, Jean-Marc and Skoglund, Jan. Lpcnet: Improving neural speech synthesis through linear prediction, 2018. URL https://arxiv.org/abs/1810.11846.

Wang, Yuxuan, Skerry-Ryan, RJ, Stanton, Daisy, Wu, Yonghui, Weiss, Ron J., Jaitly, Navdeep, Yang, Zongheng, Xiao, Ying, Chen, Zhifeng, Bengio, Samy, Le, Quoc, Agiomyrgiannakis, Yannis, Clark, Rob, and Saurous, Rif A. URL https://arxiv.org/abs/1703.10135.

Wiki. Left-behind children in china. 2022. URL https://en.wikipedia.org/wiki/Left-behind_children_in_China.