CO542 – Background Report

Reproduce and Optimize The State-of-The-Art TTS System With A Compressed Model

Fei Xie (CID:01229445)

Wednesday 20th June, 2018

1 Summary of Proposal

1.1 Background

Technology of Text-to-speech(TTS) used to translate written language into huamn speaking has revolutionised in last few years, which can be applied to several scenes such as human-equipment interface, accessibility for visually impaired media and entertainment. The classic technology were based on the concatenative TTS, where to recombine large number of speech fragments from a large database containing the recordings from a single speaker. In other words, it is hard to modify the voice once finished [21]. Another multi-stage hand-engineered pipelines stated in [20] bring a widely used tool nowadays to transform text into a compact audio representation with a synthesis method, which called vocoder [7].

However, with the boom of technology of Deep Learning, especially after the success of ImageNet[6], neural networks has been applied into TTS area to build as stated end-to-end neurons system to input the image of written words and then transformed into kinds of audios as output with different voice.

Figure 2 shows the tack of development of the main used technology at present.

Figure 1: Vocoder

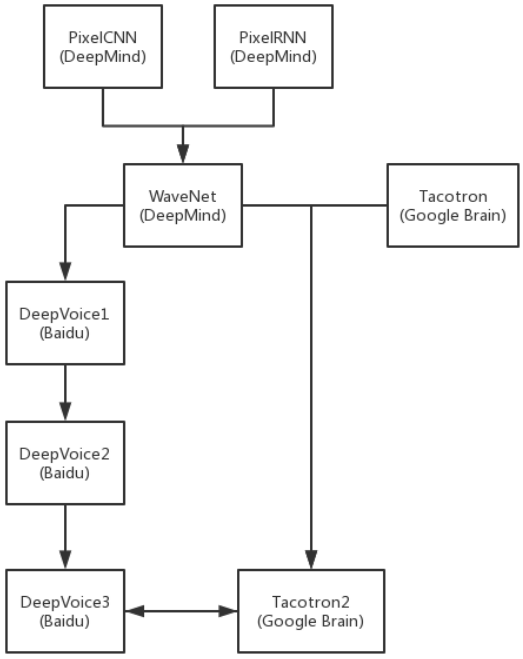


Figure 2: Development TTS System Based on Neural Networks

1.2 Analysis of Different Structure

1.2.1 Wavenet

Wavenet[21] was created by DeepMind in 2016, which was the first autoregressive model to try to model the raw audio, which adapt 16,000 samples per second as the output which showed in figure6. Additionally, these samples depend on the previous ones which inspired using the similar structure of neural network as PixelCNN[14] and PixelRNN[12] where writers stated it possible to generate complex natural images one colour-channel at a time with a structure called dilated convolution networks showed in figure 4, which can enlarge the respective fields without increasing the layers of the

network. Another structure called residual and skip connections showed in 5 was from [9] to accelerate the speed of training and ensure the gradient can be broadcast into a deeper layer.

Figure 3: Audio Samples in Different Time Scale

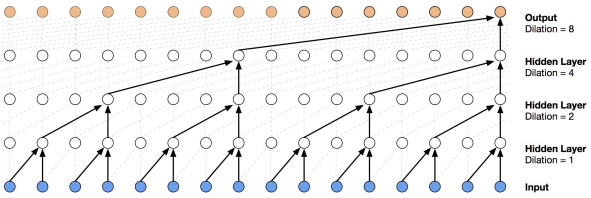


Figure 4: Structure of Dilated Convolution Networks

At the training stage, the waveforms recorded from human speakers will be the input, and the output will be the synthetic utterances after training. Then this output will be put back into the network to be a support to keep training the next value, which is implemented by the dilated networks. This process is very slow which is the main disadvantages of Wavenets [13], so DeepMind gave another more advanced network called Parallel WaveNet in[13] which use a similar structure like GANs to speed up the training process.

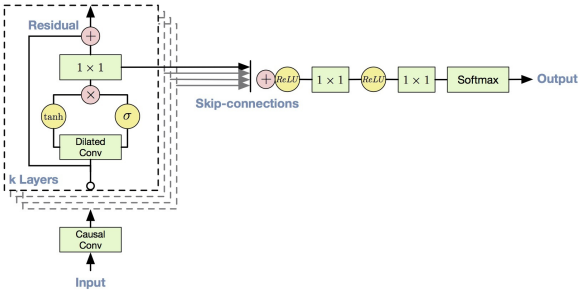
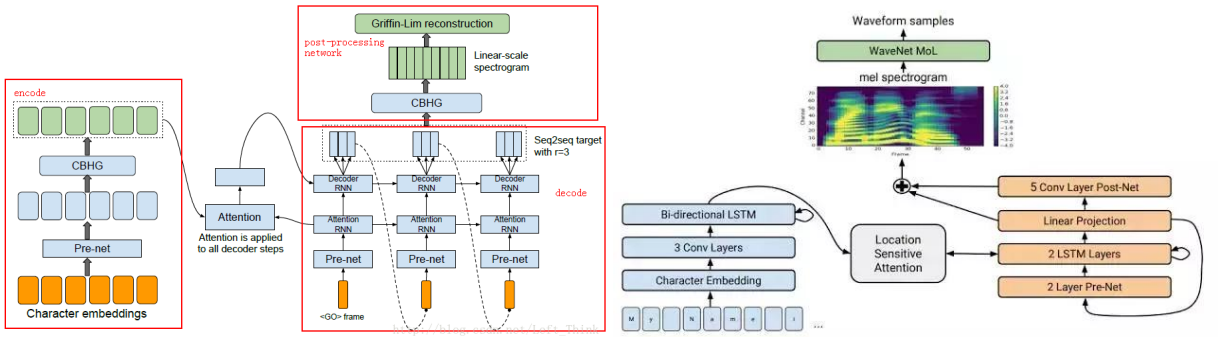


Figure 5: Residual And Skip Connections

1.2.2 Tacotron & Tacotron2

Two more structures were given by the group of Google Brain called Tacotron[22] present some other useful models to deal with TTS problems. Due to that the input to WaveNet need a lot expertise to produce, involving elaborate text-analysis systems as well as a robust lexicon, Tacotron [22] applied a sequence-to-sequence(S2S)[19] architecture, which can produce magnitude spectrograms from a sequence of characters. However, Tacotron, as showed in figure6a adapted the Griffin-Lim algorithm [8] for phase estimation which vocoded from the resulting of the S2S net. As described above, WaveNet has been proved to produce higher audio quality than Griffin-Lim algorithm and even characteristic artifacts, in which case, Tacotron2 was made to combine the advantages of Tacotron and WaveNet with a entirely neural approach to speech synthesis based on architecture of Tacotron[22] S2S model and vocoder of WaveNet [21] which was demostrated in figure6b.



(a) Structure of Tacotron (b) Structure of Tacotron2

Figure 6: Audio Samples in Different Time Scale

1.2.3 DeepVoice3

In recent years, Baidu offered several kinds of architectures of networks to solve TTS problem. Starting from Deep Voice 1[1], Deep Voice 2[2] and till now the Deep Voice 3 [15] has shown the highest performance. Deep Voice 1 2 kept the traditional structure of TTS pipelines such as separating grapheme-to-phoneme conversion at first, then perform duration and frequency prediction, and last step is to produce waveform synthesis. Compared to Deep Voice 1 and Deep Voice 2, Deep Voice 3 employs an attentionbased sequence-to-sequence model, which is similar to the architecture of Tacotron[22], yield-

ing a more compact architecture. However, Deep Voice 3 avoided Recurrent Neural Networks (RNNs) to speed up the whole training process, which also makes attention-based TTS feasible for a production TTS system with no compromise on accuracy by avoiding common attention errors.

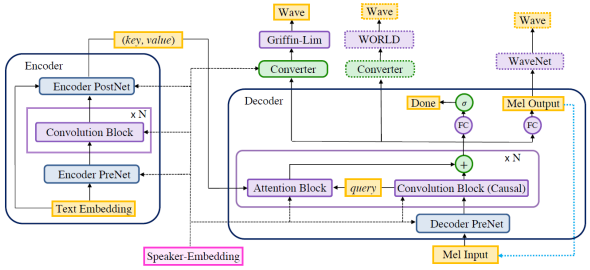


Figure 7: Structure of DeepVoice3 Networks

As shown in the figure7, the whole network contains 3 parts:

- Encoder: Here the encoder is a kind of fully-convolutional network, which converts textual features to an internal learned representation.
- Decoder: The decoder is another kind of fully-convolutional causal network, which decodes the learned representation from encoder with a convolutional attention mechanism and then transformed into a low-dimensional audio representation in an autoregressive manner.
- Converter: Converter is a fully-convolutional network as well with a post-processing structure, which is used to do prediction about final vocoder parameters (depending on the vocoder choice) from the decoder hidden states. Howerer, unlike the decoder, converter is a non-causal network so as to depend on future context information.

1.2.4 Summary

The overall aim of this research focused project is to use deep learning technology to learn hand-writting text features, and then produce corresponding audios frame-by-frame, and then try to compress the whole model so as to speed up the model with inferencing, in which case it can run with a simpler device such as mobile phone and so one. By this, it is meant that the text features can be used to estimate relative audio

features. One approach to do this is by reproducing state-of-the-art Text-to-Speech models, which were trained end-to-end to make TTS translation, which will aid greatly to develop basic deep learning skills.

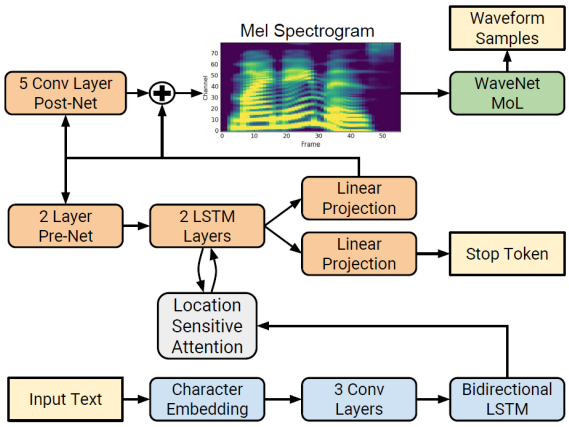
Table 1: Comparison The Performance of TTS Models

Model Name	MOS(Mean Opinion Score)
WaveNet	4.341 ± 0.051
Tactron	4.001 ± 0.087
Tactron2	4.341 ± 0.051
DeepVoice2	3.69 ± 0.23
DeepVoice3	4.331 ± 0.051
Ground truth	4.582 ± 0.053

According to the table1 shown above, DeepVoice3 and Tactron2 will be the most candidates to be reproduced and optimized. Considering the useful materials and code available online, Tacotron2 illustrated a better performance than DeepVoice3 and a simpler architecture which can be easier to be reproduced and improved. So this project will pay the main attention on Tactron2.

2 Overall General Design

Figure 8: Data Flow Diagram[18]



As shown in figure 8, the whole structure is designed to implement a end-to-end model to make the input as text, and output as according waveform. The whole struc-

ture will include two components which are

1. a recurrent S2S network used to make prediction of features with attention mechanism. The output will be a sequence of mel spectrogram frames.

2. WaveNet which has been modified with a faster and smaller architecture will be used to generate time-domain waveform samples from the output of S2S network, which is mel spectrogram.

2.1 System Components

2.1.1 Spectrogram Prediction Network

This part of network consist of three part which are an encoder, a decoder and an attention. The input of characters sequence will be converted into a hidden feature representation by encoder, then decoder will use this information to make prediction about the spectrogram. Here the input character are represented by a learning character embedding with size of 512-dimension, then being passed into a 3 stacked convolution networks each containing 512 filters with shape of 5x1. Batch Normalization [11] and ReLu function will be applied to speed up the training process. The output of the final layer will be passed into another structure of network called bi-directional[17] LSTM[10] which has 512 units to produce features in need. Then attention network will receive these features to summarize the full encoded sequence to be fixed-length vector which is the same with the size of output of decoder. It should be noted that the attention mechanism [4] used here, which adapted cumulative attention weights to add up the feature from previous decoder steps, extends the traditional one called additive attention mechanism[3]. In this condition, some sub-sequence contained within some potential failure modes will be repeated or neglected by decoder so as to promote the network moving forward consistently. After projecting inputs and location features into some hidden representations with 129-dimensional, the attention probabilities will then be calculated. For the decoder, which is an autoregressive recurrent neural network(RNN), it will be used to predict a mel spectrogram per frame at a time by the input sequence which is from encoder. Additionally, a small network called pre-net con-

taining 2 fully connected layers of 256 hidden ReLu units will process the prediction at first, then the output will be concatenated with the context vector made by attention to feed to a stack of 2 uni-directional LSTM layers. The concatenation of the LSTM output and the context vector from attention net will be projected within a liner transform to make prediction for the spectrogram frame. Finally, the made mel spectrogram will be sent into a 5-layer convolutional post-net which is used to make improvement onto the whole networks with adding to the prediction. Furthermore, each post-net layer contains 512 filters with shape of 5x1 with batch normalization.

2.1.2 WaveNet Vocoder

WaveNet is used to invert the mel spectrogram feature produced by the prediction network stated above into some time-domain waveform samples. However, due to it is slow and hard to train, which use 3 dilation cycles made from 30 dilated convolution layers, a new version which only adapt 2 upsampling layers in the conditioning stack instead of 3. In addition, PixelCNN++[16] and Parallel WaveNet [13] inspired a new idea with a 10 component mixture of logistic distribution to generate 16-bit samples at 24 kHz. The output of WaveNet is passed to a ReLU activation followed by a linear projection, in which case, to predict parameters for each mixture component. Furthermore, the loss function will be negative log-likelihood of the ground truth sample.

2.1.3 Intermediate Feature Representation

For this project, a low-level representation called melfrequency spectrograms will be chosen to be the data transformed between the two systems stated above. The reason for choosing this data type is due to that firstly it is easily computed from time-domain waveforms, which helps us train the two basic systems separately, and secondly it is smoother than waveform samples and its loss can be measured by squared error method. Here the mel-frequency spectrogram has some relationship with linear-frequency spectrogram which is also called short-time Fourier transform (STFT). With using a non-linear transform to the frequency axis of the STFT, mel-frequency spectrogram is easy to achieved. The advantages of applying this auditory frequency scale will

result in some effect which can emphasize details in lower frequencies. This is critical to speech intelligibility especially de-enohasizing high frequency details. In this situation, features derived from the mel scale has been the representation of speech recognition [5]. Compared to the data used in WaveNet, mel spectrogram is a simpler and lower level features, so it can be straightforward to be a input of a similar WaveNet model to generate audio, which has the similar function as vocoder[7].

3 Measurement

In this project, a measurement method called Mean Opinion Score(MOS), which is widely used in the domain of quality of experience and telecommunications engineering, is used to measure the performance of reproduced Tacotron2 Model. This ratings can be usually collected in some subjective quality evaluation tests, but can be algorithmically estimated as well.

Nowadays, MOS is a commonly used measurement for video, audio, audiovisual quality evaluation and so on. MOS contains a range of 1 to 5 to represent the quality of the audio, where 1 is the lowest and 5 represents the highest perceived quality. However, some other MOS ranges are also useful if the rating scale need to be scaled to a large range based on the test. In this project, the Absolute Category Rating scale is used, which maps ratings between Bad and Excellent to numbers between 1 and 5 which is shown in table2.

Table 2: Rating scales and mathematical definition

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

Within a subjective quality evaluation test, MOS will be computed as the arithmetic mean over single ratings performed by human subjects for a given stimulus.

$$MOS = \frac{\sum_{n=1}^{N} R_n}{N}$$

4 Project Timeline

4.1 Timeline Figure

TIMELINE 1: *Timeline of this Project*

15.06.2018 Finish All Background reading

01.07.2018 Understand Tacotron2 Code with Pytorch

15.07.2018 Reproduce Tacotron2 with TensorFlow

01.08.2018 Speed Up Tacotron2 with Optimization Method

15.08.2018 Compress Tacotron2 Model

25.08.2018 Try to Transfer Tacotron2 onto Mobile Device(Alternative)

05.09.2018 Finish All Report

07.09.2018 Final report submission

4.2 Specific Key Problems in Timeline

The key part of this project is to reproduce and then try to optimize the whole structure of Tacotron model. There are two ways to do this.

- Improve Algorithm: Try to remove some filters during the inference process, which is stated as compress the model, so as to reduce the network size and speed up the computing, at the same time, trying to keep the whole performance of the network unchanged.
- Engineering Optimizing: Try to make full use of the implementation libraries such as tensorRT in TensorFlow, which can accelerate the process to 1.6 times of the speed, to optimize the network with processing stage.

5 Bibliography

References

[1] Sercan O Arik et al. “Deep voice: Real-time neural text-to-speech”. In: *arXiv preprint arXiv:1702.07825* (2017).

[2] Sercan O Arık et al. “Deep voice 2: Multi-speaker neural text-to-speech”. In: *arXiv preprint arXiv:1705.08947* (2017).

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. “Neural machine translation by jointly learning to align and translate”. In: *arXiv preprint arXiv:1409.0473* (2014).

[4] Jan K Chorowski et al. “Attention-based models for speech recognition”. In: *Advances in neural information processing systems*. 2015, pp. 577–585.

[5] Steven B Davis and Paul Mermelstein. “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”. In: *Readings in speech recognition*. Elsevier, 1990, pp. 65–74.

[6] Jia Deng et al. “Imagenet: A large-scale hierarchical image database”. In: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE. 2009, pp. 248–255.

[7] Daniel W Griffin and Jae S Lim. “Multiband excitation vocoder”. In: *IEEE Transactions on acoustics, speech, and signal processing* 36.8 (1988), pp. 1223–1235.

[8] Daniel Griffin and Jae Lim. “Signal estimation from modified short-time Fourier transform”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32.2 (1984), pp. 236–243.

[9] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[10] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[11] Sergey Ioffe and Christian Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift”. In: *arXiv preprint arXiv:1502.03167* (2015).

[12] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. “Pixel recurrent neural networks”. In: *arXiv preprint arXiv:1601.06759* (2016).

[13] Aaron van den Oord et al. “Parallel WaveNet: Fast High-Fidelity Speech Synthesis”. In: *arXiv preprint arXiv:1711.10433* (2017).

[14] Aaron van den Oord et al. “Conditional image generation with pixelcnn decoders”. In: *Advances in Neural Information Processing Systems*. 2016, pp. 4790–4798.

[15] Wei Ping et al. “Deep voice 3: Scaling text-to-speech with convolutional sequence learning”. In: *Proc. 6th International Conference on Learning Representations*. 2018.

[16] Tim Salimans et al. “Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications”. In: *arXiv preprint arXiv:1701.05517* (2017).

[17] Mike Schuster and Kuldip K Paliwal. “Bidirectional recurrent neural networks”. In: *IEEE Transactions on Signal Processing* 45.11 (1997), pp. 2673–2681.

[18] Jonathan Shen et al. “Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions”. In: *arXiv preprint arXiv:1712.05884* (2017).

[19] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. “Sequence to sequence learning with neural networks”. In: *Advances in neural information processing systems*. 2014, pp. 3104–3112.

[20] Paul Taylor. *Text-to-speech synthesis*. Cambridge university press, 2009.

[21] Aaron Van Den Oord et al. “Wavenet: A generative model for raw audio”. In: *arXiv preprint arXiv:1609.03499* (2016).

[22] Yuxuan Wang et al. “Tacotron: Towards End-to-End Speech Synthesis”. In: *arXiv preprint arXiv:1703.10135* (2017).

6 Appendix:Ethics Checklist

Sections	Yes	No
Section 1: HUMAN EMBRYOS/FOETUSES		
Does your project involve Human Embryonic Stem Cells?		✓
Does your project involve the use of human embryos?		✓
Does your project involve the use of human foetal tissues / cells?		✓
Section 2: HUMANS		
Does your project involve human participants?		✓
Section 3: HUMAN CELLS / TISSUES		
Does your project involve human cells or tissues? (Other than from "Human Embryos/Foetuses" i.e. Section 1)?		✓
Section 4: PROTECTION OF PERSONAL DATA		
Does your project involve personal data collection and/or processing?	✓	
Does it involve the collection and/or processing of sensitive personal data (e.g. health, sexual lifestyle, ethnicity, political opinion, religious or philosophical conviction)?		✓
Does it involve processing of genetic information?		✓
Does it involve tracking or observation of participants? It should be noted that this issue is not limited to surveillance or localization data. It also applies to Wan data such as IP address, MACs, cookies etc.		✓
Does your project involve further processing of previously collected personal data (secondary use)? For example Does your project involve merging existing data sets?		✓
Section 5: ANIMALS		
Does your project involve animals?		✓
Section 6: DEVELOPING COUNTRIES		
Does your project involve developing countries?		✓
If your project involves low and/or lower-middle income countries, are any benefit-sharing actions planned?		✓
Could the situation in the country put the individuals taking part in the project at risk?		✓
Section 7: ENVIRONMENTAL PROTECTION AND SAFETY		
Does your project involve the use of elements that may cause harm to the environment, animals or plants?h		✓
Does your project deal with endangered fauna and/or flora /protected areas?		✓
Does your project involve the use of elements that may cause harm to humans, including project staff?		✓
Does your project involve other harmful materials or equipment, e.g. high-powered laser systems?		✓

Section 8: DUAL USE		
Does your project have the potential for military applications?		✓
Does your project have an exclusive civilian application focus?		✓
Will your project use or produce goods or information that will require export licenses in accordance with legislation on dual use items?		✓
Does your project affect current standards in military ethics – e.g., global ban on weapons of mass destruction, issues of proportionality, discrimination of combatants and accountability in drone and autonomous robotics developments, incendiary or laser weapons?		✓
Section 9: MISUSE		
Does your project have the potential for malevolent/criminal/terrorist abuse?		✓
Does your project involve information on/or the use of biological-, chemical-, nuclear/radiological-security sensitive materials and explosives, and means of their delivery?		✓
Does your project involve the development of technologies or the creation of information that could have severe negative impacts on human rights standards (e.g. privacy, stigmatization, discrimination), if misapplied?		✓
Does your project have the potential for terrorist or criminal abuse e.g. infrastructural vulnerability studies, cybersecurity related project?		✓
SECTION 10: LEGAL ISSUES		
Will your project use or produce software for which there are copyright licensing implications?		✓
Will your project use or produce goods or information for which there are data protection, or other legal implications?		✓
SECTION 11: OTHER ETHICS ISSUES		
Are there any other ethics issues that should be taken into consideration?		✓