Review Paper for TTS Algorithm

Gauri S. Nandkhedkar[1], Prajakta A. Ghumatkar[2], Vinayak Kabra[3] and Eesha Bhayya[4]

[1,2,3,4]Information Technology Department, Pune Institute of Computer Technology, Pune, India

Email: [1]gaurinandkhedkar99@gmail.com, [2]ghumatkarprajakta@gmail.com, [3]vinayakkabra25@gmail.com and [4]bhayyaeesha@gmail.com

***Abstract*—** The evolution of Text To Speech has seen many algorithms from the conventional Concatenative Synthesis to the most evolved Google's Tacotron and its iteration Tacotron 2. This survey paper discusses the architecture and outcomes of the three recent models for TTS which are Wavenet, Tacotron 1, and Tacotron 2, and further compares all of them based on their respective mean opinion scores(MOS). The MOS Values for the above-mentioned algorithms are 4.21, 3.82, and 4.58 respectively. The paper concludes that Tacotron 2 has the highest MOS value and hence is also widely used across various applications of Text to Speech.

***Keywords*—** Text-to-speech, Deep learning, Wavenet, Tacotron.

I. INTRODUCTION

This paper aims at comparing the most popular algorithms in the field of Text To Speech synthesis. This survey paper compares three TTS algorithms, namely Wavenet, Tacotron 1, and Tacotron 2. These algorithms are studied in detail and compared based on their MOS values to decide the algorithm that is best suited across various applications in the field of TTS.

II. LITERATURE SURVEY

A. Wavenet

A-1. Introduction

Wavenet[1] is a generative model that directly operates on raw audio waveforms, which means that it is not directly used to completely generate speech output from text input, instead, it is used to enhance the quality of the output by operating on the audio waveforms. The wavenet is structured as a fully convolutional neural network with these convolution layers. Wavenet can generate nearly accurate human-sounding audio output by training the models using neural networks with recordings of actual human speech. Wavenet has been tested on US English and mandarin which gave excellent results and its performance is even better than Google's best text-to-speech systems. Wavenet is in fact developed by Google DeepMind.

A-2. Architecture

Wavenet is a type of Feedforward Convolutional Neural Network (CNN) that takes raw waveform inputs and synthesizes corresponding outputs one at a time. Briefly, as discussed in the paper, this model is fed raw waveforms as input of speech in English and Mandarin which passes through the network and learns different sets of rules to describe how the audio waveform evolves. Therefore this highly trained network can then be used to create new speech-like waveforms at 16,000 samples per second[1].

This model is used to accurately model the speech output based on different parameters like the voice, accent, and tone. This makes it applicable in Music as well, which means if music waveforms are fed to it, it will produce musical output.

This model makes use of six major components to achieve the desired results. These six components are as follows:

1. Dilated Casual Convolutions: This is the major part of the model. These allow the network to operate on a coarser scale as compared to a normal convolution. Additionally, the stacked dilated casual convolutions increase the receptive field of the wavenet.
2. Softmax distributions: Since Raw audio is stored at 16-bit integer values, a softmax layer would have to output 65,536 probabilities for every single timestamp to model all possible values, which is a very large number. So we perform non-linear quantization by applying μ-law companding transformation to quantize it to 256 possible values. The results discussed in the paper proved that this reconstructed signal resulted in better output.
3. Gated Activation Units: These are used in the Gated PixelCNN which work significantly better than the rectified linear activation function.
4. Residual and Skip Connections: Both of these connections are used to speed up the process of training much deeper models.
5. Conditional Wavenets: This is required when there is an extra input to be considered. By giving extra input we guide the model to consider and produce output with the desired characteristics.
6. Context Stacks: It is only one other complementary way to increase the receptive size of a Wavenet.

A-3. Results

Wavenets produce a high Mean Opinion Score(MOS) i.e. 4.21, which is a subjective evaluation of resemblance of speech/sound produced artificially.

However, the Wavenet does not enforce long-term consistency which can result in second to second variation and deviation in instrumentation, sound quality, and voice reproduction.

B. Tacotron 1

B-1. Introduction

A text-to-speech synthesis system typically consists of multiple stages, such as a text analysis frontend, an acoustic model, and an audio synthesis module. Building these components often requires extensive domain expertise. But using Tacotron, an end-to-end generative text-to-speech model we can directly synthesize speech from characters. The model can be trained completely from scratch with random initialization.

B-2. Architecture

The foundation of the Tacotron is a seq2seq model which includes an encoder, an attention-based decoder, and a processing net. At a high level, this model takes input characters and produces spectrogram frames, which are then converted into waveforms.

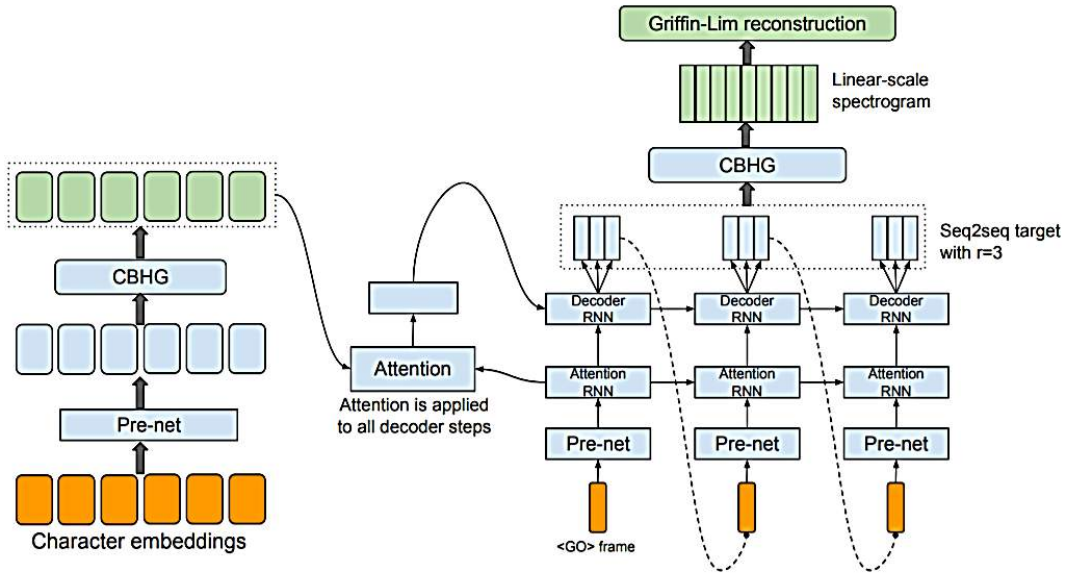

Fig.1[2] System Architecture of Tacotron 1

It consists of four important modules which are as follows:

1. CBHG module: CBHG stands for 1-D Convolutional Bank + Highway network + bidirectional GRU. It is a robust module for extracting representations from the sequences.
2. Encoder: The encoder receives an input of character sequence to extract sequential representations of the text where each character is represented as a one-hot vector and embedded into a continuous vector. A set of non-linear transformations combined and called a "pre-net", is applied to each embedding. These pre-net outputs are reconstructed by a CBHG module, into the final encoder representation which is used by the attention module.
3. Decoder: A stack of GRUs with vertical residual connections is used for the decoder. The convergence is speeded up by the residual connections. A different target for seq2seq decoding and waveform synthesis is used due to redundancy. A simple fully-connected output layer is used to predict the decoder targets.
4. Post-processing net and waveform synthesis: To convert the seq2seq target to a target that can be synthesized into waveforms is the task of the post-processing net. The post-processing net learns to predict spectral magnitude sampled on a linear-frequency scale as the synthesizer Griffin-Lim is used and it can see the full decoded sequence.

B-3. Results

Tacotron 1 achieves a mean opinion score of 3.82 on US English, with a clean and straightforward waveform synthesis module. In addition, since Tacotron generates

speech at the frame level, it's substantially faster than sample-level autoregressive methods.

C. Tacotron 2

C-1. Introduction

Tacotron 2[3] is developed by Google and the University of Berkeley in the year. In traditional TTS algorithms, the process followed includes segmentation, phonetics, and the identification of speech features, frequency, duration which are predicted using a vocoder. But in the case of Tacotron 2, it is an end-to-end speech synthesis model which is fed with character sequence input and directly predicts mel spectrogram output. This can later be used to generate audio waveforms.

C-2. Architecture

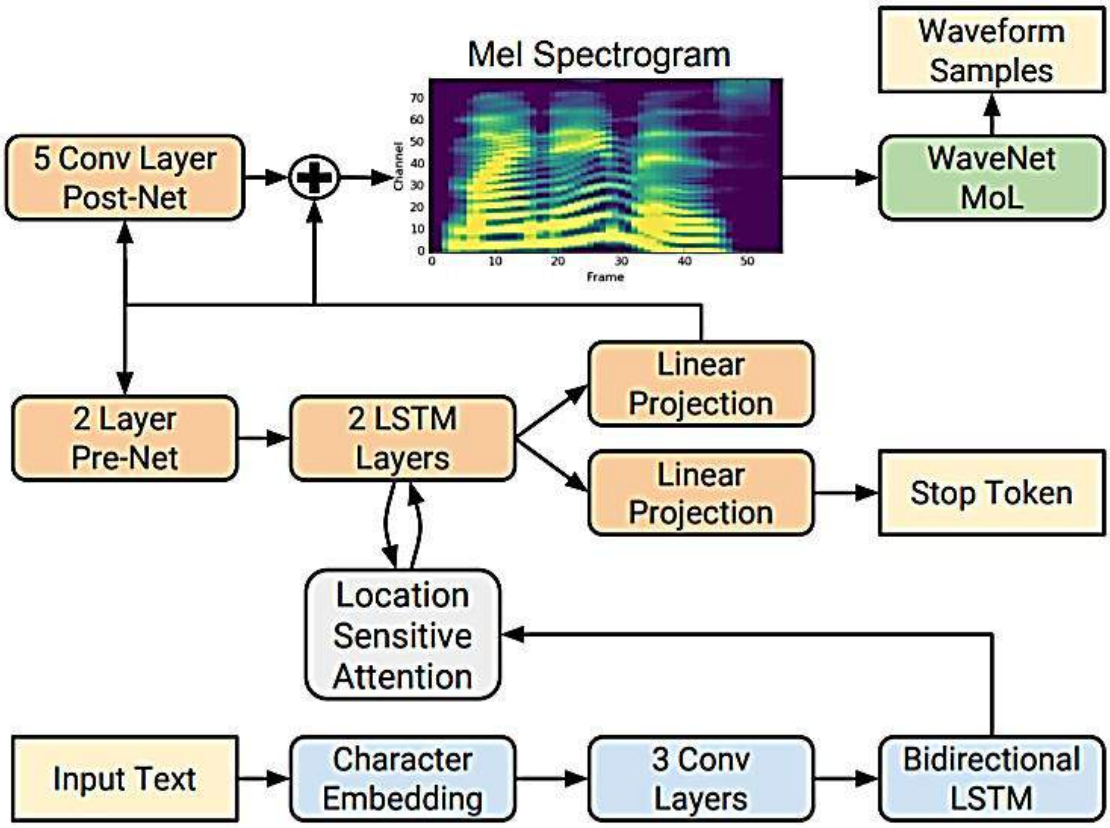

Fig.2[3] System Architecture of Tacotron 2

The components colored in blue in the diagram 2 represent the Encoder along with the input text while those colored in red form the decoder part.

1. Encoder: Firstly, character embedding is performed on the input text, during which we get a fixed dimensional vector for each character. Later we get a matrix of numbers and each vector is the embedded vector for that particular character. Next, we use three convolutional layers which capture the special relations of the character embedded matrix. It is a 1-dimensional convolution as we are dealing with a sequence model and not an image. The bi-directional LSTM is used to capture the temporal information or the temporal relation between the characters.
2. The output of the encoder is given to the Location Sensitive Attention, which is used to enhance the final results. Consider a scenario of predicting the output at a particular timestamp ‘t’, this component not only considers the previous output but also considers the current input sequence and tries to get relevant information from the relevant part and use it to predict the current output. The output of this component is fed to the decoder.
3. Decoder: Here, firstly in the 2 Layer LSTM, for every timestamp of the decoder, it predicts the mel spectrogram i.e. 40 or 80-dimensional feature vectors. The output of this is given to the Linear

projection components. These processes are carried out iteratively and the Stop Token component is used to stop this iteration. The 2 Layer Pre-net component used in this recursive model helps for the frame-to-frame prediction.

4. The output of the decoder results in the generation of the mel spectrogram. Here, this algorithm makes use of the Wavenet[1] model on the mel spectrogram, as discussed earlier in this paper, to enhance the quality of the speech that is finally generated.

C-3. Results

Trained directly on normalized character sequences and corresponding speech waveforms, this model learns to synthesize natural-sounding speech which is difficult to differentiate from the real human speech. The model achieves a mean opinion score (MOS) of 4.58 for professionally recorded speech.

III. COMPARATIVE ANALYSIS

To compare the above discussed algorithms for the TTS, we have compared their respective Mean Opinion Score (MOS) values:

Table No. 1 Comparative analysis of TTS algorithms

Sr. No.	Model	MOS
1.	Wavenet	4.21
2.	Tacotron 1	3.82
3.	Tacotron 2	4.58

We observe from table no. 1 that Tacotron 1, which is an iteration of Tacotron I achieved by applying the wavenet model on mel-spectrogram, gives the best result in terms of MOS values.

IV. CONCLUSION

As we compared the MOS values for the three models- Wavenet, Tacotron 1 and Tacotron 2, we observed that Tacotron 2 has the highest value which is 4.58. We have also discussed the implementation of Wavenet which plays an integral role in producing voice from spectrograms.

REFERENCES

[1] Oord, Aaron van den ,Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. ”Wavenet: A generative model for raw audio.” arXiv:1609.03499 (2018)

[2] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, etal., ”Tacotron: Towards end-to-end speech synthesis”, Proc. Interspeech, pp. 4006-4010, Aug. 2017.

[3] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan et al., ”Natural tts synthesis by conditioning wavenet on mel spectrogram predictions”, 2017.