ISBN: 978-1-948012-17-1

Ethics and Information Technology (ETIT)

DOI: <http://doi.org/10.26480/etit.02.2020.113.116>

REVIEW OF TEXT TO SPEECH USING DEEP CONVOLUTION

Aditya Pandya, Abhishek Bhole, Arnav Shrivastava, Mrs. Vineeta Rathore

Medi-Caps University, Indore (M.P.), India

*Corresponding Author Email: rafebarber7203@gmail.com

This is an open access article distributed under the Creative Commons Attribution License CC BY 4.0, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

ARTICLE DETAILS

Article History:

Received 25 October 2020
Accepted 26 November 2020
Available online 03 December 2020

ABSTRACT

Speech is one of the oldest and most natural medium of human interaction. So, it is understandable that there is a need today, to make our computers interact using speech as well. In recent times, recurrent neural networks (RNN) have been the pinnacle of sequential data modeling techniques. They have achieved an apparent ubiquity as far as techniques to model sequential data are concerned. However, training RNN is a time-consuming task. A typical RNN can take anywhere from a few days to a few weeks to be trained. Also, RNN requires notoriously strong hardware to be trained. The weaker the hardware, the more time it takes. According to recent studies, Convolutional Neural Networks are proved to be easier and cheaper to train due to high parallelizability. This paper aims at analyzing and reviewing a novel TTS technique using CNN instead of RNN that can alleviate the massive economic costs of training on RNN based models. In this experiment, an ordinary gaming PC was used to train the network. It took considerably less time to train than its recurrent counterpart and the quality of the resulting speech was acceptable.

KEYWORDS

TTS, MEL Spectrogram, FT (Fourier Transform), STFT (Short time FT).

1. INTRODUCTION

TTS or Text-To-Speech is a growing necessity in a world which is zooming towards a new and technologically better future. To promote the further development of TTS it is necessary to create reliable, maintainable, and flexible TTS component which is accessible to a person without a technical background, business individuals and small organizations which don't have computers capable to perform complex operations for a longer period of time. Traditional TTS systems are not very user friendly as they are often an integration of several task specific modules which are not easy to operate for a novice. For example, a traditional TTS system would comprise of a module used for text analysis, a module which generates spectrograms and another module which uses all this information to generate a waveform as an output etc. Deep learning largely helps us in uniting these building blocks into a single model or models depending on the functionality, and directly maps a relation between input and output; this is called 'end-to-end' learning. The idea is to represent the usefulness and need of a novel, concise and efficient neural TTS which is fully convolutional. The architecture of this TTS is largely similar to Tacotron but is based on a fully convolutional sequence-to-sequence learning model (Barron, 2017). We therefore aim at showing that this handy TTS works reasonably without demanding high resources. The basis of this is two-fold: Reviewing an already proposed fully CNN-based TTS system that can be trained much faster than an RNN-based state-of-the-art neural TTS system, while the sound quality is still acceptable (Efficiently Trainable, 2017).

2. PRELIMINARY

To understand how this works, you first need to know the representation of sound in the world of data science and what kind of various operations we can perform on this representation to acquire suitable outputs. Some

Basic Concepts Are Mentioned below:

2.1 Fourier Transform (FT)

Before we start with what Fourier Transform is, let us first understand some terminologies that we will need to know about to understand it in a better way (Yi-Wen, 2015).

- Time-Domain Representation or a Wave Form - A Wave Form or an audio signal is the raw representation of a sound wave as a function of Time and its Intensities. This representation is also known as the Time-Domain Representation or the Time-Amplitude Representation (Yi-Wen, 2015).

Figure 1: Time-Domain Representation or a wave form

Quick Response Code	This paper was presented at	Access this article online	
	International Conference on Contemporary Issues in Computing (ICCIC-2020) - Virtual IETE Sector V, Salt Lake, Kolkata From 25th-26th July 2020	Website: www.intelcomp-design.com	DOI: 10.26480/etit.02.2020.113.116

- **Frequency-Domain Representation** - A frequency domain representation of an audio signal shows us the frequency distribution of various amplitudes over the life time of the audio wave.

Figure 2: Frequency-Domain Representation

Simply put, a time-domain graph represents changes in signals over time whereas a frequency-domain graph represents bracket of a signal which lies within each given frequency band over a range of frequencies. Hence, it is necessary for us to understand the utility of both the domain in Signals.

Following figure (Figure 3) shows the difference between visualization of Time Domain and Frequency domain

Figure 3: Time-Domain vs. Frequency-Domain Representation

Relation of Fourier Transform to all of this:

Fourier transform is a function which takes the time-domain representation of a sound wave and converts it into a frequency-domain representation by decomposing the wave into its constituent frequencies. A musical chord can be represented using both the frequency and the constituent notes in it, this is a pretty similar analogy for the same. Essentially a Fourier Transform, when applied on a Signal Represented in the Time-Domain, gets converted to the Frequency-Domain.

Figure 4: Application of Fourier Transform

Now the question arises, why do we need to convert the signal to a Frequency-Domain representation?

Time-domain representation does not provide enough information about the sound wave for us to start performing various operations; it only tells us the value of a sound wave (function) at a particular time stamp. To gain additional information like the rate at which the signal varies we need another representation known as the Frequency-domain representation. But this does not mean we can simply ignore the Time-domain representation of the sound wave, we need information from both to fully understand the information contained in a sound wave. While the time-domain representation gives us information about the time at which the sound was perceived the frequency-domain representation tells us how the frequencies in that sound wave varies.

What Frequencies Constitutes This WaveForm ? In What Order Does These Frequencies Occur In ?

Figure 5: Time Domain and Frequency Domain

2.2 Short-Time Fourier Transform (STFT)

STFT is a function that gives the frequency and phase distribution for sections of input signals as it propagates over time (Yi-Wen, 2015). This is done by dividing a longer sound wave (Time-domain) into smaller equal length fragments and then comparing the Fourier Transform of each fragment separately. This gives us the spectrum of each fragment which is plotted as a function of time. In other words, this reveals the Fourier spectrum on each shorter segment. One then usually plots the changing spectra as a function of time.

Figure 6: Short-Time Fourier Transform (STFT)

2.3 Spectrograms

As we discussed earlier that both Time-domain and Frequency-domain representations are equally important to understand the true nature of a sound wave (Mel Spectrogram, 2019).

Using STFT the Time-domain representation of a sound wave can be converted into a 3 vectored Time-Frequency-Amplitude Representation called the spectrogram. To visualize this 3D representation we simply color-code (Heat map) the Amplitude and represent the Time and Frequency of the 2D Scale as shown below in Figure 7.

Figure 7: Working of a Spectrogram

2.4 3D- Visualization of a Spectrogram

Figure 8: 3D Visualization of a Spectrogram

2.5 MEL Scale

The MEL scale is a simple mathematical transformation of the frequency scale which is designed such as sounds that are at equal distance on the Mel scale sounds like they are equidistant from other. This needed to be done because in the Hz scale frequencies like 600 Hz and 1100 Hz are easily differentiable where as the difference between frequencies like 7600 Hz and 8100 Hz is barely noticeable (Mel Spectrogram, 2019; Oppenheim, 1970)

2.6 MEL Spectrogram

Now we know what a spectrogram and a Mel scale is, if we simply replace the frequency scale on the y-axis with the Mel scale we get a MEL Spectrogram (Oppenheim, 1970)

Figure 9: MEL Spectrogram

3. IMPLEMENTATION

The DC-TTS Model consists of 2 main Model:

3.1 Text2Mel Model

We train this model to synthesize a coarse MEL Spectrogram from a given Text. This model consists of 44 sub-modules.

- Text Encoder- Encodes the input sentence $L = [c_1, c_2, c_3, \dots, c_n]$ where $c_1, c_2, c_3, \dots, c_n$ are characters, into two Key and Value Matrices K and V
- Audio Encoder- Encodes the coarse MEL-Spectrogram of previously spoken speech, whose length is T , into a matrix Q . Takes in as input, the target audio Spectrogram and outputs a queries matrix Q of size $d \times T$, where T is the length of the audio clip.
- Attention- An Attention Matrix A evaluates how strongly the n^{th} character c_n and t^{th} frame S are related. $A_{nt} \sim 1$ implies that the modules are looking at n^{th} Character c_n at time frame t and it will look at c_n or c_{n+1} character around them, at the subsequent time frame $t+1$.
- Audio Decoder- The resultant R is concatenated with the encoded audio Q as $R' = [R, Q]$. Then R' is decoded by the Audio Decoder module to synthesize a coarse MEL-Spectrogram. The result Y is compared with the temporally-shifted truth S by a loss function and error is back propagated to the network parameters

3.2 Spectrogram Super-Resolution Model

Super Resolution is the process of up scaling and or improving the details within an image. A low-resolution image is taken as an input which is up scaled to a higher resolution, which is the output. Just like we see in the James Bond movies where an image is zoomed in and the quality of the image is enhanced magically, the magic that happens behind the scenes is called Image Super-Resolution. Similarly, our SSR model does the same but for audio signals or spectrograms.

Figure 10: Spectrogram Super-Resolution Model

4. CONCLUSION

This paper described the Deep-Convolution Text-To-Speech Model with an attention mechanism. The Deep Convolution Text-To-Speech model can be trained over a time period of around 15 hours using a standard gaming PC with almost acceptable result.

Even though the quality of the outputs generated are far from perfect but we believe better outputs can be obtained by longer training and adjusting some hyper-parameters.

The required training time and resources are both found out to be considerably lower than the previously used RNN model and thus this model opens up a lot of new possibilities and uses. The CNN model can successfully be used, for example, on mobile devices as well as on high-end computers.

The quality of the text generated is not perfect as mentioned earlier but perfection can be achieved through future research and work in the area. This method expands the use of Text to Speech model implementation to various devices and system due to its low computation and high efficiency.

ACKNOWLEDGEMENT

We would like to express our gratitude to our academic institute Mediacaps University, Indore (M.P.) for the constant support in our Bachelor of Technology (Computer Science) study and this research, for the steadiness, motivation, expertise and extensive help. Faculty's supervision helped us in making this review flourish in our knowledge and career.

REFERENCES

- Barron A., 2017. Implementation of Google's Tacotron in TensorFlow, Available at GitHub, <https://github.com/barronalex/Tacotron>
- Deep Learning and neural networks, 2017. via <https://towardsdatascience.com/introducing-deep-learning-and-neural-networks-deep-learning-for-rookies-1-bd68f9cf5883>
- Efficiently Trainable Text-to-Speech System Based on Deep Convolutional Networks with Guided Attention, 2017, <https://arxiv.org/abs/1710.08969>
- Getting to Know the Mel Spectrogram, 2019, <https://towardsdatascience.com/getting-to-know-the-mel-spectrogram-31bca3e2d9d0>
- Goodfellow I. et al., 2016, Deep Learning, MIT Press, <http://www.deeplearningbook.org>.
- Oppenheim, A.V. 1970. Speech Spectrograms using the Fast Fourier Transform. Spectrum, IEEE. 7. 57 - 62. 10.1109/MSPEC.1970.5213512.
- Yi-Wen L., 2015, The Short-Time Fourier Transform. http://ocw.nthu.edu.tw/ocw/upload/130/1603/10410%E5%8A%89%E5%A5%95%E6%96%87%E6%95%99%E6%8E%88%E6%95%B8%E4%BD%8D%E6%95%B8%E4%BD%8D%E8%81%B2%E8%A8%8A%E5%88%86%E6%9E%90%E8%88%87%E5%90%88%E6%88%90L3_Lecture-3-STFT.pdf.