RESEARCH ARTICLE

SAMRIDDHI Volume 12, Special Issue 2, 2020 Print ISSN: 2229-7111 Online ISSN: 2454–5767

Text to Speech Synthesis in Celebrity's Voice

Ajinkya P. Gaddime, Dhananjay P. Mane, Ruchita K. Vehale, Vaishnavi S. Khawale, D. G. Bhalke*

Electronics and Telecommunication Department, All India Shri Shivaji Memorial Society's College of Engineering, Pune, Maharashtra, India

Abstract

This paper is proposed for text to speech synthesis. It uses neural network architecture for generation of speech and its synthesis directly from text in celebrity's voice. The device is fitted with a recurring sequence-to-sequence prediction that graphs the embedding characters into mel scale spectrograms, followed by an updated WaveNet model that functions as a vocoder to create time-domain waveforms from those spectrograms. Here, project evaluation of the impact of mel spectrograms as the conditioning input to WaveNet rather than linguistic features, length, and F0. This paper further would be showing that utilizing this compact acoustic intermediate representation allows a significant reduction in the size of the WaveNet architecture.

Using this technique, we are going to modulate the output of the vocoder according to the frequency and pitch of a specific celebrity. Using a unit selection method of concatenation synthesis, a database of prerecorded voice is collected. This paper includes creating a database of an Indian celebrity, clustering, indexing, and synthesizing it for creating a voice output with respect to the text as input. Also worked on normalization of text which includes abbreviations, acronyms, and linguistic analysis. This paper gives output for phonemic features, like vowel length, vowel height, frontness, consonant voicing, consonant poi, and position in the syllable and word.

Keywords: Acoustic distance measure (ADM), Artificial neural network (ANN), Coevolutionary deep neural networks (CNNs), Deep neural network (DNN), Recurrent neural networks (RNNs), Text-to-speech (TTS).

SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology (2020); DOI: 10.18090/samriddhi.v12iS2.6

Introduction

Generating natural speech from text remains a demanding task despite decades of study. Over time, different techniques have subjugated the field. Statistical speech synthesis, which directly generates smooth trajectories of speech features to be synthesized by a vocoder, followed by solving many of the problems that concatenative synthesis has its periphery artifacts. The audio created by these devices, however, often sounds muffled and unnatural compared to human discourse.

WaveNet is a generative model of waveforms of the time domain, generating audio quality that starts to equal that of human speech and is used in some text-to-speech (TTS). The inputs to WaveNet, however, require significant domain expertise to produce, involving elaborate text-analysis systems, as well as, a robust pronunciation guide. Tacotron, a sequence-to-sequence architecture for producing magnitude spectrograms simplifies the conventional speech synthesis pipeline from a sequence of characters by replacing the output of certain linguistic and acoustic features with only one neural network trained from complete data alone as shown in block diagram Figure 1. For vocoding the resulting magnitude spectrograms, Tacotron uses the Griffin-Lim algorithm for phase estimation and is followed by an inverse short-time Fourier transform as Griffin-Lim produces characteristic artifacts and lower audio fidelity than approaches like WaveNet.

Corresponding Author: Dr. D. G. Bhalke, All India Shri Shivaji Memorial Society's College Of Engineering Pune, India, e-mail: dgbhalke@aissmscoe.com

How to cite his article: Gaddime, A.P., Mane, D.P., Vehale, R.K., Khawale, V.S., Bhalke, D.G. & Itkarkar, R. R. (2020). Text to Speech Synthesis in Celebrity's Voice. *SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology,* 12(SI-2), 27-30.

Source of support: Nil

Conflict of interest: None

Basics of TTS

Synthesis of speech is the artificial production of human discourse. A computer device that is used for this purpose is called a speech computer or speech synthesizer, which may be found in software or hardware items. A TTS system translates natural language text into speech; other systems

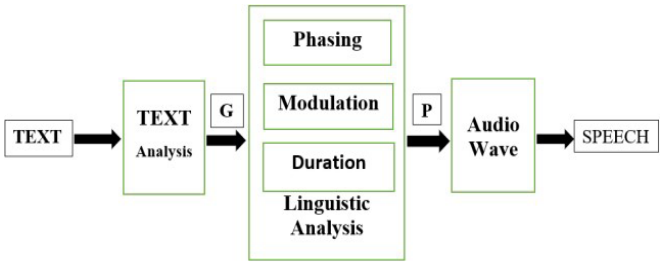

Figure 1: Text-to-speech (TTS)

turn abstract linguistic representations into speech, such as, phonetic transcriptions. Synthesized speech can be generated by concatenating captured speech pieces, which are stored in database systems vary in the size of the stored voice units; a system that stores phones or diphones has the greatest output range, but may lack clarity. Storing whole words or sentences for specific usage domains allows for high-quality output. Alternatively, a synthesizer may get incorrect. A TTS (or "engine") system consists of two parts: a front-end, and a back-end.

There are two foremost tasks to the front-end. First, it converts raw text with symbols, such as, numbers and abbreviations into the equivalent of written-out words. This process is often called standardization of the text, pre-processing, or tokenization. The front-end then assigns to each word phonetic transcriptions, separating and marking the text into prosodic units, such as, phrases, clauses, and sentences. The method of assigning phonetic transcriptions to words is called conversion either from text to phoneme, or from grapheme to phoneme, or from grapheme to phoneme. The symbolic linguistic representation which is produced by the front-end is made up of phonetic transcriptions and prosody information. The rear-end is often called the synthesizer.

Deep Neural Networking

Deep learning (also known as deep structured learning or hierarchical learning) is part of a larger, artificial neural network-based family of machine learning methods. Training may be directed, semi-supervised, or unmonitored.

Deep learning architecture, such as, deep neural networks, deep belief networks, recurrent neural networks, and convolutional neural networks have been practiced in this field, including computer vision, speech recognition, natural language processing, voice recognition, social network scanning, machine translation, bioinformatics, drug design, medical image analysis, and device inspection. A deep neural network (DNN) is an artificial neural network (ANN) multilayered between the input and output layers. The DNN determines the exact mathematical manipulation to transform the input into output, whether it is a linear relation or a nonlinear relationship. The network moves by measuring the likelihood of each production through the layers. For example, a DNN trained to recognise dog breeds will go over the given image and measure the likelihood that the dog in the image is of a certain breed. One can review the results and select the probabilities that should be displayed by the network and return the projected label. Mathematical manipulation as such is called a layer, and there are numerous layers in complex DNN, hence, the term "deep" networks. Typically, DNNs are feed forward networks where data transfer takes place from the input layer to the output layer without looping backwards. Firstly, DNN creates a graphical view of virtual neurons and assigns the connections between them to random numerical values, or "weights." Multiply the weights and inputs and return an output between 0 and 1. If a particular pattern was not correctly recognized by the network, an algorithm would change the weights. This helps the algorithm make those parameters more effective until the required mathematical procedure to process the data in its entirety is determined. Recurrent neural networks are used in applications, such as, language modeling, in which data can gush in any direction. For this purpose, long short-term memory is especially effective.

In computer vision, coevolutionary deep neural networks (CNNs) are used. CNNs have also been applied to automated speech recognition (ASR) acoustic modeling. Tacotron is an end-to-end, text-to-speech generative system that synthesizes speech directly from text input. Through random initialization, it can be fully equipped from scratch. It generates speech at the frame level faster than sample-level autoregressive methods.

Methodology

There are essentially three stages involved in the text to the speech conversion process, which are referred to as text to words, words to phonemes, and phonemes to sound.

Text to Word

Reading words is easy, but if one listens to a young child reading a book that was just too hard for them, one can understand it's not as trivial as it seems. The main difficulty is that written text is ambiguous, the same written information can have more than one meaning and usually one has to understand the meaning or make an educated estimate to read it correctly. So, the initial stage of speech synthesis, which is commonly called normalization, is all about eliminating ambiguity; it is about narrowing down the many probable ways you might read a piece of text into one that is the most suitable.

Word to Phonemes

Phonemes are to spoken language what letters are to written language, they are the fragments of spoken sound—the sound factors from which one can make any spoken word. The word kite is made up of four "k" (as in kangaroo), "i" (as in pit), "t" (as in tusk), and "e" (as in eat) phonemes. Rearrange

the order of the phonemes and may render the words "act" or "tack." In the English alphabetic order, there are just 26 letters but more than 40 phonemes. That is because few letters and letter groups can be read in several ways (a, for example, can be read differently, as in 'pad' or 'paid'), so instead of one phoneme per letter, there exist phonemes for all the different letter sounds. Some languages require more or less phonemes than others typically in the range of 20–60.

Phoneme to Sound

Completing the conversion of text into a list of phonemes, now making the basic phonemes that the computer reads out loud when it is turning text into speech. There are three different methods for this. Firstly, to use human recordings saying the phonemes. Secondly, for the computer to produce the phonemes itself through the generation of basic sound frequencies, and the last one is to mimic the human voice mechanism.

System Design

Training to Model

Intermediate Future Representation

We will be choosing a low-level acoustic representation, mel frequency spectrograms, to bridge the two components of the system. The use of an easily computed representation from time-domain waveforms permits us to train the two components individually. Also, this representation is smoother than waveform samples and is simpler to train using a mean squared error loss because phasing within each frame is invariant. It is obtained by using a nonlinear transformation to the short-time Fourier transform (STFT) frequency axis, influenced by measured reactions from the human auditory system, and summarizes the lesser-dimensional frequency contents (Figure 2).

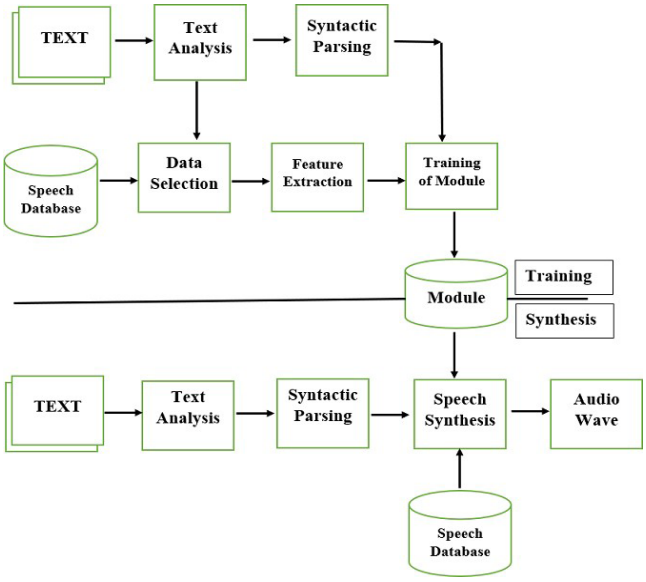

Figure 2: Block diagram

Spectrogram Prediction Network

In the Tacotron model, mel spectrograms are calculated using a frame size of 50 ms using an short-time Fourier transform (STFT) as shown in Figure 3. Transformation of the STFT magnitude to the mel-scale using an 80 channel mel filter bank. The network is attentively having an encoder and a decoder. The encoder transforms a sequence of characters into a representation of a hidden function that the decoder consumes to predict a spectrogram. The characters that are given as input are represented using 512-dimensional character embeddings which are bypassed through a stack of 3 convolutional layers, each containing 512 filters. Each filter spans 5 characters, followed by batch normalization and rectified linear unit (ReLU) activations. As in text to speech converter, these convolutional layers model longer-term context in the incoming character sequence.

WaveNet Vocoder

WaveNet is a deep convolutional artificial neural network. It is perfectly suited to solving various complex problems in speech processing. It is also a fundamentally new and amazingly powerful method for implementing voice transformation, speech amplification, and speech compression. Its structure is a network that auto-completes an occluded image, according to its content by generating pixel predictions from a pixel's nearest neighbors.

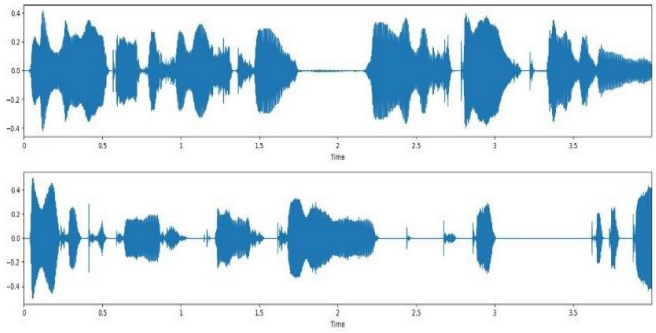

Figure 3: Spectrogram

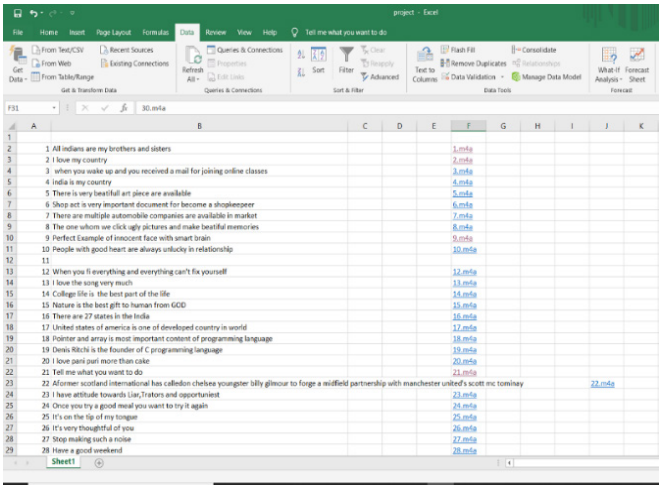

Figure 4: Recorded audio database

Voice Database

We are first trying to develop this module in our own voice with a pre-recorded audio database as shown in Figure 4. By referring to this database, our module gets established by concatenation and unit selection methodology.

Output

Speaker: Text to speech conversion; the voice processing module transforms .txt file to an audio output using unit selection and concatenation methodology via speaker in the desired celebrity's voice.

Result

Synthesized voice of an Indian celebrity by creating a database of pre-recorded speech and gained the output in the celebrity's voice with text input for the clustered database of 500 sentences.

Conclusion

This paper describes TTS synthesis in a celebrity's voice, a fully neural TTS system that combines a sequence-to-sequence recurrent network with attention to predict mel spectrograms with a modified WaveNet vocoder. The resulting device synthesizes speech with audio quality at the Tacotron level, both prosody, and WaveNet. This system is trained for generating output voice of celebrity's frequency and pitch directly from text data, without relying on complex feature engineering and reaches the state-of-the-art sound quality near to that of normal human speech.

Acknowledgment

The authors take this prospect to express their deep sense of gratitude towards their guide Prof. Dr. D. G. Bhalke, under whose guidance they had the privilege to work on this project. The authors also convey their warm gratitude towards their project co-coordinator Prof. R. R. Itkarkar, for her kind support and co-operation. The authors are grateful to Dr. D. G. Bhalke, Head of E and TC Engineering Department, and all teaching and non-teaching staff members of E and TC Engineering Department, for their direct or indirect help in the completion of this task.

References

[1] Jonathan Shen1, Ruoming Pang1 and Ron J. Weiss (2018), Natural TTS synthesis by conditioning Wave-Net on Mel spectrogram predictions, University of California, Berkeley,.

[2] Aaronvanden Oord, Sander Die leman and Heiga Zen (2016), Wave-Net: A Generative model for raw audio, Google DeepMind, London, UK.

[3] Tom Le Paine, Pooya Khorrami, Shiyu Chang, Yang Zhang (2016), Fast Wave-Net generation model, University of Illinois at USA.

[4] P. Taylor (2009), Text-to-Speech Synthesis, Cambridge University Press, New York, NY, USA, 1st edition,.

[5] N. Swetha and K. Anuradha (2013), Text to speech conversion, International Journal of Advanced Trends in Computer Science and Engineering, Vol .2, No.6, Pages : 269-278

[6] H. Zen, A. Senior, and M. Schuster (2013), "Statistical parametric speech synthesis using deep neural networks," in Proceedings of ICASSP, pp. 7962–7966.