An Intelligent Approach of Text-To-Speech Synthesizers for English and Sinhala Languages

Pabasara Jayawardhana[1]
Achala Aponso[2]
Naomi Krishnarajah[4]
Informatics Institute of Technology,
Colombo, Sri Lanka
e-mail: pabasarajayawardhana@gmail.com[1],
achala.a@iit.ac.lk[2], Naomi@iit.ac.lk[4]

Amila Rathnayake[3]
University of Colombo School of Computing,
Colombo, Sri Lanka
e-mail: amilamad@gmail.com[3]

***Abstract*—This paper attempts to investigate novel Text-to-Speech algorithm based on Deep voice which is an attention based, fully convolutional mechanism. The procedure of producing speech synthesis involves with learning statistical model of the human vocal production mechanism which is eligible of taking some text and vocalize that as speech. This paper would reveal the route of the attempt where there is the destination of accuracy and realism. Serenity and fluency are the most important qualities which expect from a TTS. The idea is to give an outline of discourse amalgamation in the Sinhala language, compresses and replicates about the characteristics of different blend procedures utilized. The proposed TTS synthesizing with the neural network based approach to perform phonetic-to-acoustic mapping has described by the purpose of applying for multilingual synthesizers.**

Keywords-text-to-speech; speech synthesis; natural language processing; ANN; CNN

I. INTRODUCTION

The quality of TTS, basically the quality of speech synthesizer is decided by its similarity to the human voice and by its ability to be understood clearly. The absence of expressions in the sound for a TTS makes a huge impact on the usage of the application. Hence, it declares the major problem of a TTS development, which is the generation of the sound out of a text which is natural as Human voice. The target of the TTS is to be able to create the full range of human speech which includes all the speech variations and reduce the robotic gasp of the output voice to enhance the gap between human performance and machine voice. Sinhala is the mother tongue of most of the Sri Lankans and it is also one of the official and national languages of Sri Lanka. Since there are many people in Sri Lanka who use Sinhala to communicate, there is a need to pay attention to the research area of recognizing Sinhala speech. Sinhala language is extremely varied, too complex and subtle for computers to understand and replicate. Hence there are only few attempts taken to build a TTS for Sinhala language and still it exists as a major research area to investigate is one of the motivations of this research.

II. BACKGROUND

Stephen Hawking, the brilliant scientists were only able to convey his brilliant ideas to the world thanks to the advancement of digital technology. He was suffering with Amyotrophic Lateral Sclerosis (ALS) disorder and he needed smart, assertive technologies to communicate with others. Text-to-Speech is one of assertive technologies which help to variety of impaired people such as Stephen to improve their interaction with the real world. The main purpose of TTS or Read digital text into speech applications is reading the text captured from anywhere into the voice by maintaining similar voice to a human. Nowadays, TTS can be considered as a technology which leverage the human-computer interactions as well as web accessibility helps visually and vocally impaired people in their communications.

The best answer to the question whether we really want a TTS to sound like a human at all and wouldn't they be happy with a "robotic" sounding one instead, which would be much easier to produce. The most listeners are extremely intolerant of unnaturalness to the extent that they will refuse to use non-natural sounding systems regardless of what other benefits are provided. There is a significant difference between "listening" and "hearing." According to the Rose & Dalton[1] , "Hearing seems effortless, automatic and nonselective …On the other hand, listening feels intentional; it is effortful, focused and selective. We need to be awoken to listen. Hearing is reactive, while listening is strategic" [1]. As per the above discussion to have a proper listening and understanding, TTS should be accurate that is needed to achieve human performance, which has a more natural voice with accents which is understandable.

Most of the speech recognition systems that have been developed so far are for English speech recognition than other languages [2] with different speech synthesis methods. For English language, there are many TTS applications which becoming popular day by day and most of them still under research and development to improve the performance of these systems. However, this state-of-the-art on TTS conversion is limited of converting text of few languages and still subsists as an active topic where many researches carried out trying to build technology for the conversion of text to

978-1-7281-3323-2/19/$31.00 ©2019 IEEE

Authorized licensed use limited to: Informatics Institute of Technology. Downloaded on May 22,2020 at 16:11:03 UTC from IEEE Xplore. Restrictions apply.

speech models identical to human capability. On the other hand, for Sinhala language still, there is no any TTS application, which could help to leverage the computer literacy of the Sri Lankan community.

The major complication of building a TTS application is that, it must learn so many language rules which are difficult to say as a fixed set of rules. People can say the same word differently. And one combination of a letter can sound differently in two places. Also, there are regional differences in words so called accents or dialects. That is, people in the same region speak the same word in different forms which differ in tone and accent. Moreover, men speak in a different tone than women [3]. Sometimes sound contains emotions as well. Even the available TTS still struggle to achieve human performance, which could capture all these variations. Due to the difficulty of making a model which can capture the dialects and required tone to add to the text most text to speech programs speak text without any accent or tone which reads a collection of words. Some applications use various voices as the output and these voices can speak in many of the most common accents, and pronounce words, according to the rules for different languages.

On the other hand, in terms of machine learning, still TTS conversion exists as one of a problem in natural language processing (NLP) which is trying to solve by many researchers hence it is more beneficial to many domains. And yet there is no any machine learning succeeded approach in language invariant speech generation. And like other speech synthesizers, NLP models also suffocate with the common problem of eliminating robotic voice.

III. Sophisticated Neural Network Based Approaches for TTS

WaveNet is a deep neural network designed for generating raw audio waveforms [4]. WaveNet model creates raw audio waveforms from scratch. The stack of outstanding Convolutional layers without connections in between are the structure introduced in Wave Net and has speed up the generation of models by using dilated convolutions. This model has conditioned by feeding vocoder parameters from a TTS which is pre-existing by the WaveNet researchers. The result was mind-blowing and it produced high quality voice which was extremely clean with no noise. The main practical problem with this model is the expensiveness, for the computation in normal conditions. WaveNet uses 40 of convolution layers including other connections to generate High-quality speech. It is generating 1 second of 16kHz audios involves preparing 16000 samples at a time [5]. This isn't an achievable speech synthesis system, at any rate not with the present innovation, and the assets that a normal developer can bear Since that it has investigated with different techniques.

Merlin is designed to model statistical parametric speech synthesis from a Deep Neural Network .There is a front end text processor and a vocoder with the Merlin implementation [6]. The system trains a neural network to predict acoustic features and after that process it send to a vocoder to produce the speech waveform. Merlin is an open source toolkit which is written using python construct on theano library. Merlin uses currently available acoustic models as Festival or Ossian as its front end. It uses the standard feedforward neural network, long short-term memory (LSTM) recurrent neural network, mixture density neural network and recurrent neural network (RNN) as the targeted neural networks. The target idea of this implementation is to compare different type of neural networks and benchmark the ideals and open a new research area on neural network base speech synthesis

Deep Voice is introduced with the traditional text-to-speech pipelines and receives a similar structure, while replacing all parts with neural systems and utilizing less complex features. It's a fully-convolutional architecture for speech synthesis, which is constructed with attention-based mechanism. Architecture of Deep Voice 3 consists with three components as Encoder, Decoder and Converter. Inside learned representation is proceeding with a multi-hop convolutional attention mechanism in an autoregressive manner .While comparing with the recently published attention-based text to speech systems as Tacotron, Deep Voice 3 has ten-fold increase in training speed [7].This model has the capability of handling multi-speaker speech synthesis successfully. Since Deep Voice 3 describes common error approaches in sequence-to-sequence models and show the successful path to avoid those errors, the experimental design was based on the Deep voice.

IV. The Need for A New Toolkit

Most of the speech recognition systems that have been developed so far are for English speech recognition than other languages [8] with different speech synthesis methods. For English language, there are many TTS applications which becoming popular day by day and most of them still under research and development to improve the performance of these systems. However, this state-of-the-art on TTS conversion is limited of converting text of few languages and still subsists as an active topic where many researches carried out trying to build technology for the conversion of text to speech models matching human capability. On the other hand, there is no any TTS application for Sinhala language which could help to power the computer literacy of the Sri Lankan community.

The major obstacle to building a TTS application is to introduce a fixed set of rules since the language rules are different. People can say the same word differently. And one combination of a letter can sound different in two existences.

Also, there are regional differences in words so called accents or dialects. People in the same region speak the same word in different forms which differ in tone and accent. Moreover, men speak in a different tone than women. Sometimes sound contains emotions as well. Even the available Text to speech systems still struggle to achieve human performance, which could capture all these variations. Due to the difficulty of making a model which can capture the dialects and required tone to add to the text most text to speech programs speak text without any accent or tone which reads a collection of words. Some applications use various voices as the output and these voices can speak in many of the most common accents, and pronounce words, according to the rules for different languages.

Authorized licensed use limited to: Informatics Institute of Technology. Downloaded on May 22,2020 at 16:11:03 UTC from IEEE Xplore. Restrictions apply.

On the other hand, in terms of machine learning, still TTS conversion exists as one of a problem in NLP which is trying to solve by many researchers hence it is more beneficial to many domains. And yet there is no any machine learning succeeded approach in language invariant speech generation. And like other speech synthesizers, NLP models also suffocate with the common problem of eliminating robotic voice.

V. THE DESIGN AND IMPLEMENTATION

The proposed text to speech application architecture is an investigation based on the findings of literature reviews as the Fig 1. The evaluations based on the conclusions of system design guide to keep the requirements stable. There are four major architectural components highlighted.

A. Data Preprocessing (Text and Audio)

The characters with spacing and punctuations are the raw text. The target of the preprocessing is to feed them with appropriate performance. The main issue with those utterances were the mispronunciation some rare words, skipping words and repeat words. The preprocessing model provided a better solution to overcome the issue by normalizing the input text.

All the characters were converted to uppercase and removed all intermediate punctuation marks in the input text. Then author added a question mark or a period for the end of each utterance. The spaces of input text which introduced by the speaker, were replaced with special separator characters. Author used four different word separators for, indicating words, Relaxed pronunciation, words, the words with short pauses between and for space characters and standard pronunciation words. Audio must not have long silences, must be aligned correctly with audio for better performance.

The sliced audios of the proper format are taken from some public domains as per the Table I.

TABLE I. PREPARING AUDIOS

	English	Sinhala
Name	"LJ "Speech Dataset	"Delowak Atarin Eha" Short story
Audio resource from	Public domain	"Lisn" Sinhala Audio book
Speaker type	Single speaker	Single speaker
File format	16-bit PCM WAV Sample rate of 22050 Hz. .wav file	16-bit PCM WAV Sample rate of 22050 Hz. .wav file
Audio count	13,100	2120
Single audio duration	1 second to 10 seconds	1 second to 10 seconds
Text resource from	Same Public domain	Converted Sinhala font
		FMAbaya->Unicode ->Singlish

B. Encorder

The target user input for the system is a text and the purpose of the encoder is to convert those textual features to an internal representation of learning outcomes from the fully convolutional layers. The type of inputs which allows the by the encoder are phoneme and phoneme stress embedding rather than character embedding. Since, encoder network translates phonemes or characters into representations of vectors which can be trained.

The proportion of embedding until the target proportion is fully-connected via a chain of convolution blocks. As the next phase, the attention key vectors are created by anticipating back to the embedding proportions. The idea behind the key vectors are to calculated attention weights from each attention block. The last context vector is calculated as typical weighted vectors over the value vectors. The Encoder consists with a convolutional block which has the qualities of outstanding connection to the input, sequential gated components as not sequential with some scaling aspects. Since the sequence requires the length, inputs are lengthened with timestamps.

C. Decoder

The purpose of the decoder is to decode the learned outcomes in an auto-regressive method. Basically, decoder predicts future audios from previous audio files. The decoder is a fully convolutional and the autoregressive emphasize that it uses completely causal convolutions. The technique behind the process is translating the attention mechanism of multi-hop convolutional into an audio representation of low-dimensional (Mel-band spectrograms). The audios are handled as groups and it creates a considerable impact on the performance since decoding numerous audio collection is better than single audio.

The structure of decoder network involves with rectified linear unit (ReLU), attention blocks and output layers. ReLU is to rectifier nonlinearities with multiple group of fully-connected layers and the attention blocks consists as a series. The output layers are fully-connected and it predicts the next set of audios with a binary wrap of last frame. The dropout technique is applied to the all the fully-connected layers occur before the attention blocks excluding the first one. The predictions for the lost functionality is generated using the output spectrograms and the cross-entropy loss function is generated using "done" prediction.

The hidden position of the decoder is used by the dot-product attention mechanism and it used the per-timestep vectors from the encoder. The output vector calculated as average, weighted. The author has trained the audio data set with a single speaker, encoding position rate with one to the decoder and fixed to the encoder.

D. Converter

The purpose of the converter is predicted final output features of the post-processing network. The prediction is

Authorized licensed use limited to: Informatics Institute of Technology. Downloaded on May 22,2020 at 16:11:03 UTC from IEEE Xplore. Restrictions apply.

depending on the method which used to synthesise the waveform from the hidden states of the decoder. Converter also a fully-convolutional and rely on impending context information. Decoder outputs from its last hidden layer, are the input activations of converter network. Converter applies non-causal and non-autoregressive convolution blocks to predict vocoders parameters and use impending decoder context to forecast outputs. The author used Griffin-Lim algorithm to convert spectrograms and the loss function calculated based on that vocoder. Used loss function based on L1 loss on a linear - scale (log-magnitude) spectrograms because the authors used Griffin-Lim algorithm.

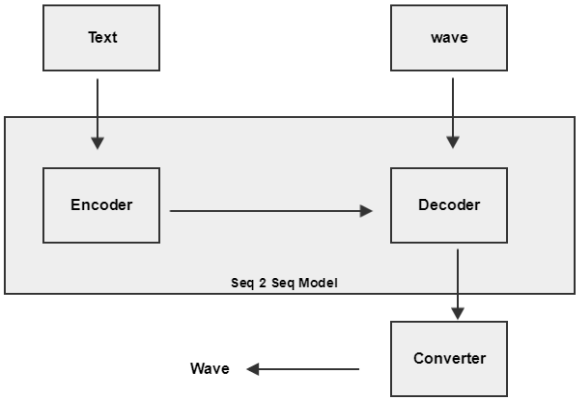

Figure 1." Architecture of the system.

VI." EXPERIMENTAL DESIGN

A. *Building Sinhala Front-End*

Text data preprocessing is a key step of a text synthesizer front end and its expectation is to have a good performance. The Sinhala front-end make sure that the text inputs were cleaned and ready for molding.

The target of building front-end is not only to clean the input text, but also to increase the training accuracy. It's consist with a method to do mix pronunciation. The words in the data set are sliced into the characters and the mix pronunciation method retrieve the relevant pronunciation from an arpabet. Author comes up with a Sinhala arpabet by creating a Sinhala phoneme database. Phonetic dictionary is a huge corpus of the Sinhala dialect, in in spoken form as well as in writing form. The input data set feed the text to text for sequence method as well and it covers the data to a number sequence.

B. *The Training Sequence to Sequence Modal*

Training model is one of the main phases of sequence to sequence speech synthesis. Training model has a preset which consists with a configuration file. The preset defines the iteration count for train by using the checkpoint interval parameter. The eval interval parameter evaluates the TTS wave form and generate the speech for giving text within the added time. Then the batch size decided by dividing total data from the iterations (No of iterations are decided from dataset / batch size). Total data refer to (no of pairs of wave and text) wave inputs and the text inputs. The number of workers represents the count of threads which use of load data the epochs parameter defines the count of global iterations and the train process set up to stop when it reaches to the relevant epochs count.

The train model feeds the input data set by keeping it with separate three objects as Text Data Source, the Linear Spec Data Source and Mel Spec Data Source. The target idea is to load training data for that pre-defined iterations. Then at the next step it builds a model by keeping the values of hyper parameter. The training happens using GPU if it is available.

C. *Synthesizing the Trained Modal*

The synthesizing model is the phase of generating a waveform for the given text from the trained model. It also creates the sequence of outcomes using checkpoints while training. An external vocoder synthesizes the trained model using the checkpoints.

D. *Subjective Evaluation*

The main achievement of a stable software system is to have the considerable accuracy with it. The accuracy is depending on the correctness or the reliability of the system and the accuracy testing is the measure point of the correctness of the system. The formula to track the accuracy testing as bellow as in (1).

$$r = \frac{x}{y} \times 100 \tag{1}$$

r = Accuracy as a percentage of the result
x = Number of input Text
y = Number of correct speech output

The conducted test approaches resulting 40% of accuracy as the current state and can improve the state by train the model with more data. The Author will bring the improve action to achieve the accuracy perfection before the completions. Finding the minimum data set as per the TABLE B and C from English speech set to adopt for the Sinhala language done using multiple training steps.

TABLE II." FINDING MINIMUM DATA SET OF ENGLISH DATA

Data Sets	Hours of training	Steps
13100	144	200 000
6500	96	100 000
2500	72	100 000
1000	24	100 000

TABLE III." FINDING MINIMUM DATA SET OF SINHALA DATA

Data Sets	Hours of training	Steps
2120	96	120 000
1000	96	100 000

Authorized licensed use limited to: Informatics Institute of Technology. Downloaded on May 22,2020 at 16:11:03 UTC from IEEE Xplore. Restrictions apply.

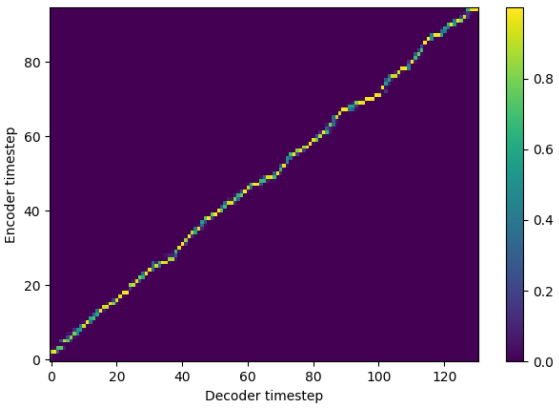

Figure 2." TensorBoard result of English wave output.

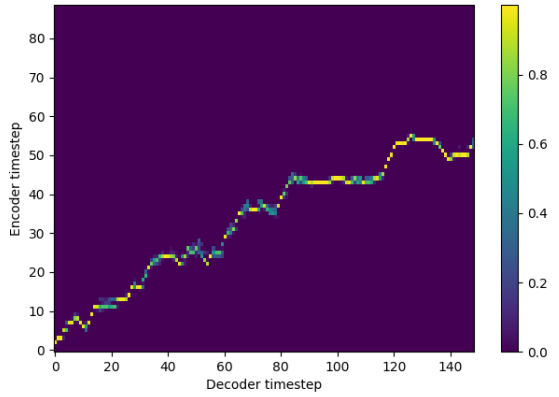

Figure 3." TensorBoard result of Sinhala wave output.

Voice quality is typically expressed in terms of the mean opinion score (MOS). Since the results are under experimental level the quality was calculated with the sense of audible range by allowing 10 users to hear the naturalness of the output audios.

VII." DISCUSSION

As per the research findings, there is no any exhibit a method of learning and mastery from open source projects which can be concentrated and effectively gotten for Sinhala language. To achieve the best outcome the line in Fig 2 should be straight. The graph with the Sinhala outcome needs more improvement since the line becomes a curve as per the Fig 3. The need of training more dataset resolves that problem. Although the existing components as speech synthesizer and text preprocessors support for English language, the lack of multilingual approaches could overcome by implementing a Sinhala support TTS. The Furthermore, as defined in the Literature review, the intelligent TTS suggests making improvements to current algorithms and approaches to add more value to the user.

This project mainly focuses on implementing an intelligent Sinhala based multilingual TTS. The solution was constructed as a machine learning approach was used neural networks. The project had huge successful and unsuccessful phases where the author required to do more novel ideas, investigations and developments to accomplish the goal which is not an easy as expected to complete. By conducting this research based project author has gained a huge amount of knowledge in a wide range in every part of a project, same as the working in an enterprise environment. Therefore, the research and project approved out as an academic project for the aim of helping academic purpose, the product has the ability and as well as the worth of being a lucrative project.

VIII." CONCLUSION

In this paper attempt to investigate that implementing a Text-to-Speech system is more effective to its users if it can generate more human speech output Among the huge number of languages which are spoken around the world there few languages which are supported by speech synthesis and among those it may, less quantity of work is done on Sinhala Text-to-Speech synthesizing. According to the Findings there are numerous methodology and strategy to produce a TTS.

The findings revel that neural network based approach to speech synthesis provides the benefits of natural speech sounding, language transferability and low storage prerequisites. The conclusions from the experiments specify that neural network based text-to-speech systems have the possibility to distribute better voice quality than traditional approaches, though some enhancement of the system is still required.

IX." FURTHER WORK

Adding phoneme dictionary to achieve more accuracy of the grapheme-to-phoneme model for the Sinhala Data set is the one of main future improvement. Other than that, to gain the appropriate accuracy the modal need to train with huge set of data as per the next step. Currently the system is working only in English and Sinhala languages. Since the basic idea of the system is to develop the TTS with multiple languages which includes in a one system, the author is having the idea to enhance the system to multilingual. As another enhancement aspect Autor is suggesting the system to integrate with multiple e-learning materials and subsystems.

X." ACKNOWLEDGMENT

We would like to express our appreciation for the all who have supported me from their efforts opinion and even by words to complete the final year research project.

REFERENCES

[1]" D. Rose and B. Dalton, *Plato Revisited: Learning through Listening in the Digital World*. 2007.

[2]" M. Peissner, "What the relationship between correct recognition rates and usability measures can tell us about the quality of a speech application," in *Proceedings of the 6th International Scientific Conference on Work With Display Units*, 2002, pp. 296–298.

[3]" L. Monaghan, J. E. Goodman, and J. M. Robinson, *A Cultural Approach to Interpersonal Communication: Essential Readings*. John Wiley & Sons, 2012.

Authorized licensed use limited to: Informatics Institute of Technology. Downloaded on May 22,2020 at 16:11:03 UTC from IEEE Xplore. Restrictions apply.

[4] A. van den Oord *et al.*, "WaveNet: A Generative Model for Raw Audio," *ArXiv160903499 Cs*, Sep. 2016.

[5] S. Mehri *et al.*, "SampleRNN: An unconditional end-to-end neural audio generation model," *ArXiv Prepr. ArXiv161207837*, 2016.

[6] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," *Proc SSW Sunnyvale USA*, 2016.

[7] W. Ping *et al.*, "Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning," *ArXiv171007654 Cs Eess*, Oct.

[8] M. Peissner, "What the relationship between correct recognition rates and usability measures can tell us about the quality of a speech application," in Proceedings of the 6th International Scientific Conference on Work With Display Units, 2002, pp. 296–298.

Authorized licensed use limited to: Informatics Institute of Technology. Downloaded on May 22,2020 at 16:11:03 UTC from IEEE Xplore. Restrictions apply.