Digital Receipt

This receipt acknowledges that Turnitin received your paper. Below you will find the receipt information regarding your submission.

The first page of your submissions is displayed below.

Submission author:	Ye Kyaw Thu
Assignment title:	NLP Paper
Submission title:	myTTS-ver3
File name:	nlp-proj.pdf
File size:	200.7K
Page count:	4
Word count:	2,080
Character count:	11,160
Submission date:	28-Jan-2021 06:03PM (UTC+0630)
Submission ID:	1496175436

Myanmar Text to Speech

Badounmar, Hnin Yu Hlaing, Hlaing May Tin,
Nan Yu Hlaing, Thida San, Zun Hlaing Moe

***Abstract*— Text-to-speech system typically consists of a text analysis front-end, an acoustic model and a speech synthesizer. Since these components are trained independently and rely on extensive domain expertise. In this project to address these problems, we apply Tacotron2 through tensorflow, which is an end-to-end generated text-to-speech (TTS) model with syllable and word-level. Given <text, audio> pairs, the model can be trained from scratch with random initialization. End-to-end system can be trained on a small number of manually labeled text and audio paired data sets brings many advantages. In our experiment, according to the difficulties of the use of GPU, we trained 10 text and speech on step 10k for a week and investigated on closed test. Tacotron2 achieves a 3.8 subjective 5-scale mean opinion score on closed tests with listeners, outperforming a production parametric system in terms of naturalness. The clarity performance of the syllable-level is better than the word-level. In addition, since Tacotron2 generates speech at the frame level, it's substantially faster than sample-level autoregressive methods.**

***Index Terms*—TTS, Tacotron, Tensorflow, MOS, Encoder, Decoder**

I. Introduction

THE main motivation for this project is to investigate the end-to-end text to speech synthesis with small corpus. TTS is a common research area of digital signal processing (DSP) and natural language processing (NLP). It is intended to generate human-like speech from the input text or sentences, a natural-sounding in terms of intelligibility and quality. The main task of TTS is to convert any text information into standard and smooth speech in real time. The speech synthesis is not a new problem, but it is still one of the challenges for organizations and businesses. The modern TTS trend is more complex. Deep learning (DL) is a new research direction in the machine learning area in recent years. In this project, we experimented with TTS by Tacotron, one of deep learning methods which is to synthesize speech directly from the characters. It does not need phoneme-level alignment and can be trained on completely from scratch given <text, audio> pairs. There are a number of generative models already exist for this purpose, but some of them are not necessarily end-to-end as they usually have models developed and trained separately. Among these models, we used Tacotron as it is truly an end-to-end generative model that can fulfill our goal.

The remainder of this paper is organized as follows. In Section 2, we describe the related work. Section 3 briefly introduces Myanmar Language. Section 4 describes methodology, Section 5 presents the overview of experimental setup, results and discussion. Lastly, we conclude in section 6.

II. Related Work

In recent years, DNN based generative models for Myanmar Speech synthesis can yield better synthesized speech than HMM [1]. In [2], Quang Pham Huu proposed a deep learning architecture to the problem of speech synthesis, Tacotron model. The output of Tactron on both BigCorpus and SmallCorpus achieved high-quality speech audio. Lwin et al., proposed Tacotron-2 model synthesize speech directly from the characters and they experimented on Myanmar text and audio pairs [3]. Kim et al., researched a generative flow of speech synthesis with monotonic alignment search without any external aligner and they obtained the comparable speech quality to Tacotron-2 [4]. Chuxiong Zhang et al., introduced Tacotron for Mandrain Chinese TTS with prosodic features to generate more natural speech and obtained better by adding the prosodic system as the front-endsystem for Tacotron [6].

III. Myanmar Language

Myanmar language is the official language of Myanmar, and it is spoken as the first language by 32 million people and as the second language by another 10 million people. Myanmar script has 33 basic consonants, 4 basic medials, 12 basic vowels, other symbols and special characters. The consonants have only 23 distinct pronunciations because some consonants have the same pronunciation in the Myanmar language. A syllable is composed of one or more characters and one or more syllables can be formed as the word in Myanmar language. If the syllable final glottal stop is regarded as a tonal feature and the non-final neural vowel as anatonic vowel, there are four phonological tones in Myanmar [5].

Badounmar is with the Faculty of Computer Science, Computer University (Thaton), Thaton, Myanmar.

Hnin Yu Hlaing is with the Faculty of Information Science, University of Computer Studies(Meiktila), Meiktila, Myanmar.

Hlaing May Tin is with the Faculty of Computer System and Technology, Myanmar Institute of Information Technology (MIIT), Mandalay, Myanmar.

Nan Yu Hlaing, Zun Hlaing Moe and Thida San are with the Faculty of Information Science, Myanmar Institute of Information Technology (MIIT), Mandalay, Myanmar.