بسم الله الرحمن الرحيم

MONA MAGHRABY

شبكة المعلومات الجامعية

التوثيق الالكتروني والميكروفيلم

MONA MAGHRABY

شبكة المعلومات الجامعية

التوثيق الإلكتروني والميكروفيلم

جامعة عين شمس

التوثيق الإلكتروني والميكروفيلم

قسم

نقسم بالله العظيم أن المادة التي تم توثيقها وتسجيلها
علي هذه الأقراص المدمجة قد أعدت دون أية تغيرات

يجب أن

تحفظ هذه الأقراص المدمجة بعيدا عن الغبار

MONA MAGHRABY

AIN SHAMS UNIVERSITY
FACULTY OF ENGINEERING
Computer and systems Engineering

Text-to-Speech Method Optimization

A Thesis submitted in fulfillment of the requirements of
Master of Science degree in Electrical Engineering
Computer and Systems Engineering Department

by

Fady Khalaf Fahmy Hakeem

Bachelor of Science of Electrical Engineering
Computer and Systems Engineering Department
Faculty of Engineering, Ain Shams University, 2017

Supervised By

Prof. Hazem Abbas

Professor of Computer and Systems Engineering Department, Faculty of Engineering, Ain Shams University.

Prof. Mahmoud Khalil

Professor of Computer and Systems Engineering Department, Faculty of Engineering, Ain Shams University.

Cairo, 2020

AIN SHAMS UNIVERSITY
FACULTY OF ENGINEERING
Computer Engineering and Systems

Text to speech Method Optimization

by

Fady Khalaf Fahmy Hakeem
Bachelor of Science in Electrical Engineering
Computer Engineering and Systems
Faculty of Engineering, Ain Shams University, 2017

Examiners' Committee

Name and affiliation	**Signature**
Prof. Mohsen Abdel Razek Rashwan Electronics and Electrical Communications Engineering Faculty of Engineering, Cairo University.	.....
Prof. Hossam Eldeen Hassan AbdElmeniem Computer Engineering and Systems Faculty of Engineering, Ain Shams University.	.....
Prof. Mahmoud Ibrahim Khalil Computer Engineering and Systems Faculty of Engineering, Ain Shams University.	.....

Date: dd mm 2020

Statement

This thesis is submitted in fulfillment of Master of Science in Electrical Engineering, Faculty of Engineering, Ain shams University.

The author carried out the work included in this thesis, and no part of it has been submitted for a degree or a qualification at any other scientific entity.

Fady Khalaf Fahmy Hakeem

Signature

.....

Date: 01 01 2020

Researcher Data

Name: Fady Khalaf Fahmy Hakeem
Date of Birth: 22/01/1994
Place of Birth: Cairo, Egypt
Last academic degree: Bachelor of Science of Electrical Engineering
Field of specialization: Computer and Systems Engineering Department
University issued the degree : Ain Shams University
Date of issued degree : 07/09/2017
Current job : Software Engineer at Mentor graphics Egypt.

Abstract

Speech synthesis is the artificial production of human speech. A typical text-to-speech (TTS) system converts a language text into a waveform. There exist many English (TTS) systems that produce mature, natural, and human-like speech synthesizers. In contrast, other languages, including Arabic, have not been considered until recently. Existing Arabic speech synthesis solutions are slow, of low quality, and the naturalness of synthesized speech is inferior to the English synthesizers. They also lack essential speech key factors such as intonation, stress, and rhythm. Different works were proposed to solve those issues, including the use of concatenative methods such as unit selection or parametric methods. However, they required a lot of laborious work and domain expertise. Another reason for such poor performance of Arabic speech synthesizers is the lack of speech corpora, unlike English that has many publicly available corpora[1] [2] and audiobooks.

End-to-end speech synthesis methods managed to achieve nearly natural and human-like speech. they are prone to some synthesis errors such as missing or repeating words, or incomplete synthesis. We may argue that this is mainly due to the local information preference between teacher forcing input and the learned acoustic features of a conditional autoregressive model. The local information preference prevents the model from depending on text input when predicting acoustic feature which contributes to synthesis errors during inference time. In this work, we compare between two

[1] LjSpeech, `https://keithito.com/LJ-Speech-Dataset/`

[2] Blizzard 2012, `http://www.cstr.ed.ac.uk/projects/blizzard/2012/phase_one/`

modified architectures based on Tacotron2[3]. The first architecture is similar to Tacotron2 but replaces the WaveNet[4] vocoder[5] with a flow-based implementation of WaveGlow[6]. It takes diacritic Arabic character sequence as input and produces mel-spectrogram per each training sample. The second architecture maximizes the mutual information between conditional text input and predicted acoustic features (mel-spectrogram) to eliminate local information preference issue. It also changes the training objective of the model by adding a Connectionist Temporal Classification[7] (CTC) loss term. Training objective could be considered as a metric of maximization of local information preference between conditional text input and predicted acoustic features. We carried the experiments on Nawar Halabi's dataset[8] which contains about 2.41 hours of Arabic speech. Our experiments show how to generate high quality, natural, and human-like Arabic speech using an end-to-end neural deep network architecture. This work uses just ⟨ text, audio ⟩ pairs with a relatively small amount of recorded audio samples with a total of 2.41 hours. It illustrates how to use English character embedding despite using diacritic Arabic characters as input and how to preprocess these audio samples to achieve the best results. It describes also how to make small changes in the existing model to achieve better results.

Keywords — Arabic text-to-speech, speech synthesis, Tacotron 2, WaveGlow, InfoGan, deep learning, neural networks

[3]Tacotron2 is a famous end-to-end TTS architecture containing 2 main parts (a) an encoder-decoder architecture with attention, and (b) a neural vocoder to synthesize speech `https://arxiv.org/pdf/1712.05884.pdf`

[4]A deep neural generative model of raw audio waveforms

[5]voice codec that analyzes and synthesizes the human voice signal for audio data compression, multiplexing, voice encryption or voice transformation

[6]A flow-based Generative neural network for speech synthesis`https://arxiv.org/pdf/1811.00002.pdf`

[7]a valuable operation to tackle sequence problems where timing is variable, like Speech and Handwriting recognition

[8]`http://en.arabicspeechcorpus.com/`

Thesis Summary

Summary

This thesis explores how to generate high-quality Arabic speech using pre-trained English model as well as pre-trained English character embedding. We utilized a transfer learning approach because the used dataset was relatively small (2.41 hours of Arabic speech). This work compares between two modified architectures based on Tacotron2[9]. The first one explores how to generate Arabic speech using pre-trained English model, while the second one further enhances the subjective quality of generated speech. Quantitative and qualitative measures have been conducted to compare between the two architectures.

This thesis is divided into six chapters, along with a list of figures, a list of tables, a list of abbreviations, a list of symbols, and a bibliography. Here is the entire structure of the thesis:

- Chapter 1 gives a quick overview of why speech synthesis is useful in real-world scenarios.
- Chapter 2 provides a literature review on speech synthesis. It covers the history of speech synthesis before and after the introduction of Deep learning approaches and end-to-end architectures.
- Chapter 3 focuses on the theoretical background that the work is based upon. It presents a quick overview of convolutional neural networks, recurrent neural networks, and long-short term memories (LSTMs). It

[9]Tacotron2 is a famous end-to-end TTS architecture containing 2 main parts (a) an encoder-decoder architecture with attention, and (b) a neural vocoder to synthesize speech `https://arxiv.org/pdf/1712.05884.pdf`