Introducing TTS System — Tacotron2

Leyuan Sheng
April 11, 2019

Table of contents

1. Introduction
2. Architecture
3. Results
4. Reference

Introduction

- A NN architecture for speech synthesis directly from text
- A recurrent Seq2Seq feature prediction network
- A modified WaveNet model acting as a vocoder

Result: sound quality close to natural human speech

Architecture

- Intermediate Feature Representation
- Prediction Network
- WaveNet Vocoder

Intermediate Feature Representation

A low-level acoustic representation: mel frequency spectrograms.

- That is easily computed from time-domain waveforms
- That is easier to train using a squared error loss because it is invariant to phase within each frame

Prediction Network

Encoder

The encoder converts a character sequence into a hidden feature representation

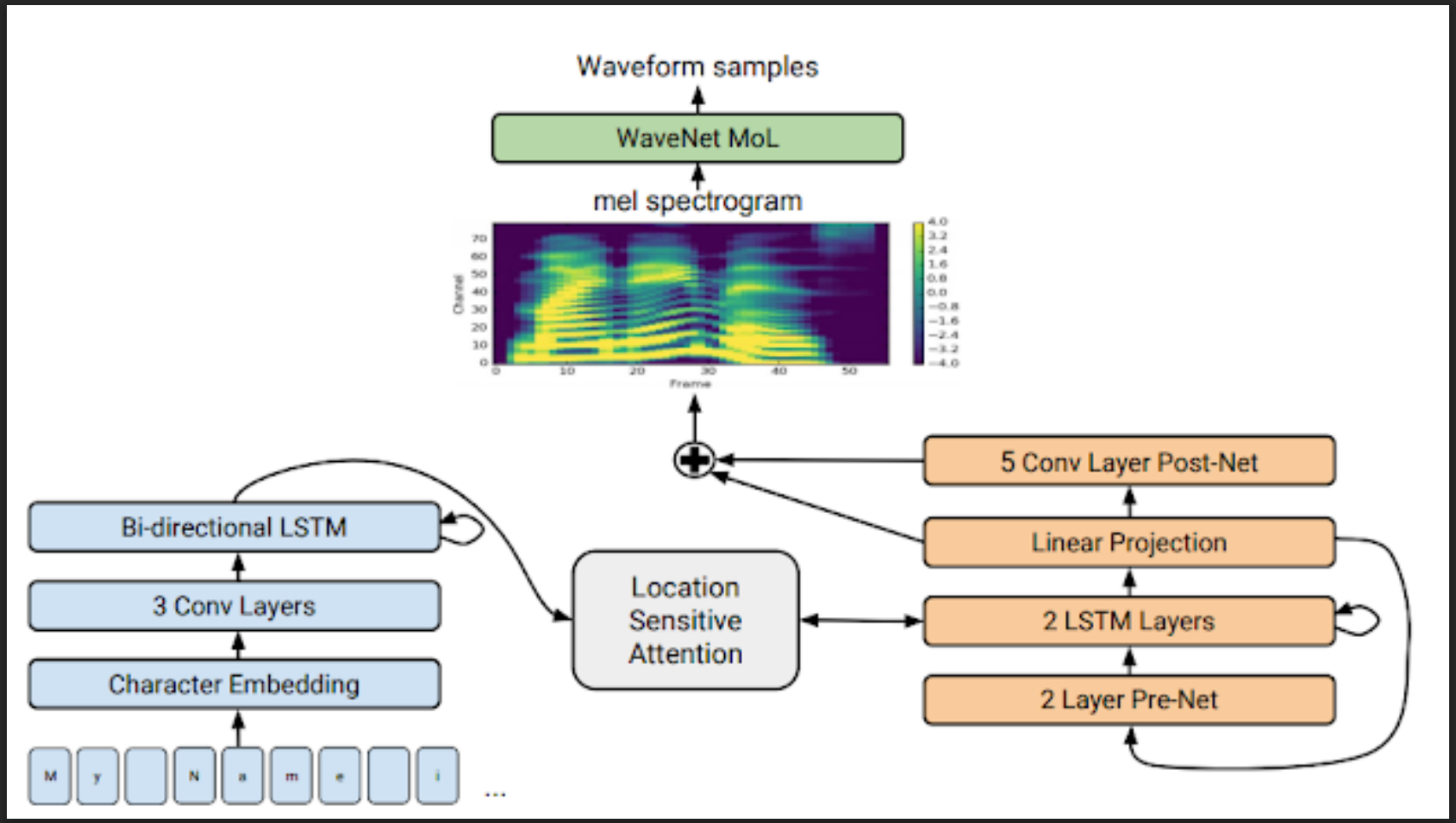
Decoder

The decoder consumes to predict a spectrogram

WaveNet Vocoder

Invert the mel spectrogram feature representation into time-domain waveform samples.

Tacotron2 system architecture



Mean Opinion Score

Table: Mean Opinion Scores

System	**MOS**
Parametric	3.492 ± 0.096
Concatenative	4.166 ± 0.091
Tacotron 2	4.526 ± 0.066
Ground truth	4.582 ± 0.053

Reference

J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, et al.
Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions.
in *ICASSP. IEEE*, 2018, pp.4779–4783.