Speech Understanding - Text-to-Speech

TTS

Subodh Kant (M23CSA531) and SyamKrishnan Sakthidharan(M23CSA535)

Department of Computer Science Engineering
Indian Institute of Technology (IIT), Jodhpur

January 31, 2025

Overview

Introduction

- This Project details the implementation and execution of a Tacotron2-based Text-to-Speech (TTS) pipeline using PyTorch and torchaudio.
- The pipeline converts text input into speech through multiple stages, including text encoding, spec-trogram generation, and vocoder-based waveform synthesis.
- The SOTA models evaluated in this study are:
  1. Tacotron2 (Spectrogram Generation)
  2. WaveRNN Vocoder (Waveform Generation)
  3. Griffin-Lim Vocoder (Waveform Generation)
  4. Waveglow Vocoder (Waveform Generation)

Text-to-Speech Pipeline

The following figure illustrates the whole process.

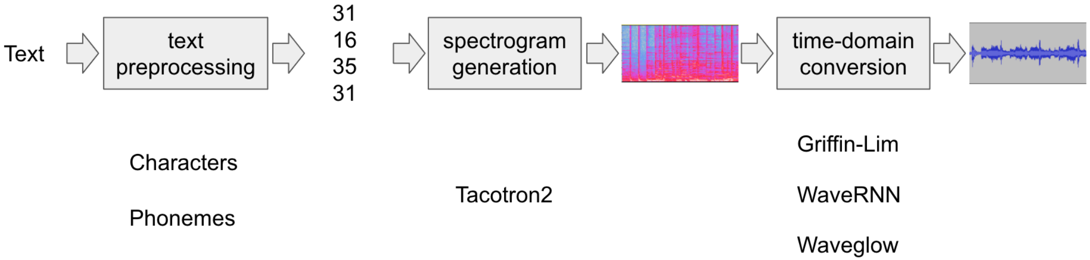

Figure: Text-to-Speech Pipeline

Text Processing

We have used below text

"The implementation and execution of a Tacotron2-based Text-to-Speech (TTS) pipeline using PyTorch and torchaudio."

Character-based encoding

The pre-trained Tacotron2 model is designed to work with a specific set of symbols (letters, punctuation, and special characters). These symbols must be converted into numeric IDs that the model can process.

```
tensor([[31, 19, 16, 11, 20, 24, 27, 23, 16, 24, 16, 25, 31, 12, 31, 20, 26, 25,
         11, 12, 25, 15, 11, 16, 35, 16, 14, 32, 31, 20, 26, 25, 11, 26, 17, 11,
         12, 11, 31, 12, 14, 26, 31, 29, 26, 25,  1, 13, 12, 30, 16, 15, 11, 31,
         16, 35, 31,  1, 31, 26,  1, 30, 27, 16, 16, 14, 19, 11,  4, 31, 31, 30,
          5, 11, 27, 20, 27, 16, 23, 20, 25, 16, 11, 32, 30, 20, 25, 18, 11, 27,
         36, 31, 26, 29, 14, 19, 11, 12, 25, 15, 11, 31, 26, 29, 14, 19, 12, 32,
         15, 20, 26,  7]])
tensor([112], dtype=torch.int32)
```

Figure: Text Processing - Output

Spectrogram Generation

Tacotron2.infer method

- Tacotron2 is the model we use to generate spectrogram from the encoded text. torchaudio.pipelines. Tacotron2.infer method perfoms multinomial sampling.

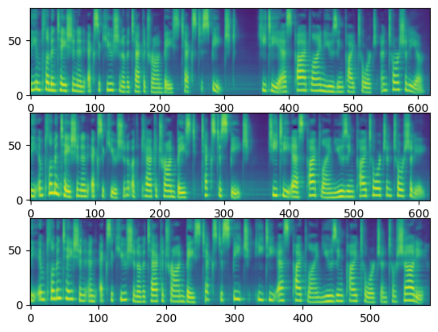

Figure: Spectrogram

Waveform Generation

- WaveRNN Vocoder:

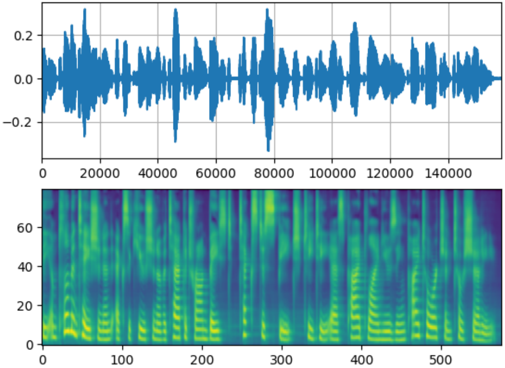

Figure: WaveRNN Vocoder

Waveform Generation

- Griffin-Lim Vocoder:

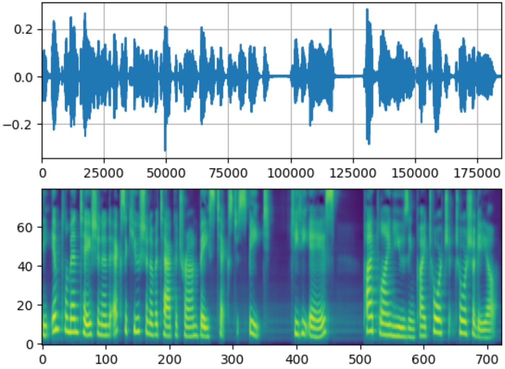

Figure: Griffin-Lim Vocodern

Waveform Generation

- Waveglow Vocoder:

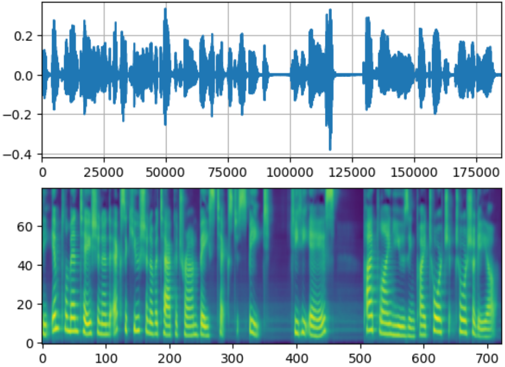

Figure: Waveglow Vocoder

Evaluation Metrics

- Mean Opinion Score (MOS): MOS is a subjective evaluation metric where human listeners rate the quality of synthesized speech on a scale of 1 (bad) to 5 (excellent).
- Results:
  1. WaveGlow and WaveRNN tend to achieve higher MOS scores (4.0-4.6) compared to Griffin-Lim (which typically scores below 3.0 due to robotic quality).
  2. Tacotron2 + WaveGlow usually achieves the highest MOS due to high fidelity and natural prosody.
- Conclusion
  1. WaveGlow + Tacotron2 performs best across MOS metrics, providing high MOS scores.
  2. WaveRNN + Tacotron2 is slightly behind but remains a good trade-off between efficiency and quality.
  3. Griffin-Lim + Tacotron2 is significantly worse in terms of perceived naturalness and spectral accuracy.

The End