NATURAL TTS SYNTHESIS BY CONDITIONING WAVENET ON MEL SPECTROGRAM PREDICTIONS

https://arxiv.org/pdf/1712.05884

0. Introduction

- 기존 TTS(Text-to-Speech) 시스템은 텍스트 분석, 발음 변환, 음향 모델, 보코더 등 여러 모듈로 구성된 복잡한 파이프라인 구조를 가짐.
- 이러한 분리된 구조는 각 단계 간의 feature mismatch 문제를 야기하고, 결과적으로 음질 저하를 유발함.
- 본 연구에서는 텍스트 입력만으로 자연스러운 음성을 생성하는 end-to-end TTS 모델을 제안함.
- 제안 모델은 Tacotron (텍스트 → 멜 스펙트로그램)과 WaveNet (멜 스펙트로그램 → 오디오 파형)을 결합하여 고품질 음성을 생성함.
- 주요 기여:
  - 완전한 신경망 기반 TTS 파이프라인 제시
  - 기존 TTS보다 자연스러운 발음, 억양, 음색을 재현
  - MOS 4.53 (5점 만점)으로 인간 음성과 거의 동일한 품질 달성

1. Overview

- 전체 모델은 두 단계 구조로 이루어짐:
  1. Text → Mel Spectrogram: 문자 입력을 받아 음성의 시간-주파수 스펙트럼(mel spectrogram)을 예측 (Tacotron 기반)

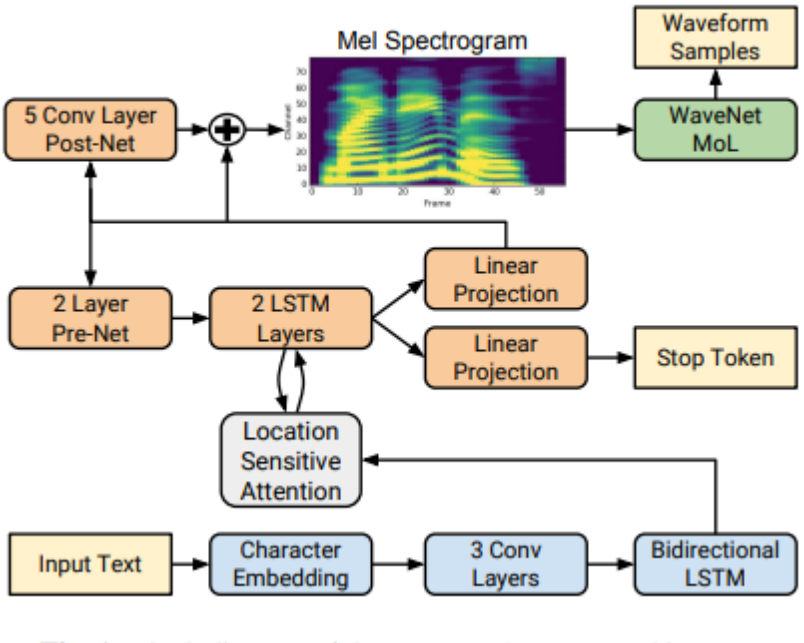
2. Mel Spectrogram → Waveform: 예측된 스펙트로그램을 WaveNet vocoder에 입력해 실제 음성 파형을 생성

- Tacotron은 sequence-to-sequence attention 모델, WaveNet은 autoregressive dilated convolution 모델.
- 두 네트워크를 분리 학습하여 안정적 학습과 높은 합성 품질을 모두 확보.
- 결과적으로 텍스트에서 오디오까지 end-to-end 파이프라인을 완성함.

2. Challenges

- Feature mismatch 문제: 기존 모듈식 TTS에서는 단계별 오류 누적 발생.
- Alignment 불안정: 텍스트-오디오 매핑에서 attention collapse 문제 발생 가능.
- 음질 저하: 기존 보코더(예: Griffin-Lim) 기반 합성은 artifacts가 많음.
- 계산량 과다: WaveNet은 샘플 단위(22kHz 이상) autoregressive 생성으로 매우 느림.
- 다화자 확장성 부족: 단일 화자 학습에 한정되어 있음.

3. Method



- Tacotron 2 Text-to-Spectrogram Network
  - Input: 문자 시퀀스
  - Encoder: Convolution Bank + Highway Network + Bidirectional GRU (CBHG)
  - Attention: Location-sensitive attention을 사용해 텍스트-음성 alignment 학습
  - Decoder: 80차원 mel-spectrogram 프레임을 autoregressively 예측
- WaveNet Vocoder
  - mel-spectrogram을 conditioning input으로 사용
  - Dilated causal convolution으로 waveform 샘플을 순차적으로 생성
  - μ-law quantization(8-bit) 사용하여 오디오 샘플링
- Training Process
  - Tacotron과 WaveNet을 별도로 학습
  - Tacotron의 ground-truth mel-spectrogram을 WaveNet의 입력으로 사용 (teacher forcing)
  - 학습 완료 후 Tacotron의 예측 스펙트로그램을 WaveNet에 전달하여 최종 합성 수행

4. Experiments / Data

- Dataset:
  - 미국 영어 단일 여성 화자 데이터, 약 24.6시간 분량
  - Sampling rate: 22.05 kHz
- Training:
  - Optimizer: Adam (lr = 1e-3)
  - Batch size: 32
  - Griffin-Lim 보코더와 WaveNet 보코더 비교
- Baselines:
  - Concatenative TTS

- Parametric TTS
- Tacotron (with Griffin-Lim)

5. Results

System	MOS
Parametric	3.492 ± 0.096
Tacotron (Griffin-Lim)	4.001 ± 0.087
Concatenative	4.166 ± 0.091
WaveNet (Linguistic)	4.341 ± 0.051
Ground truth	4.582 ± 0.053
Tacotron 2 (this paper)	$\mathbf{4.526 \pm 0.066}$

Total layers	Num cycles	Dilation cycle size	Receptive field (samples / ms)	MOS
30	3	10	6,139 / 255.8	4.526 ± 0.066
24	4	6	505 / 21.0	4.547 ± 0.056
12	2	6	253 / 10.5	4.481 ± 0.059
30	30	1	61 / 2.5	3.930 ± 0.076

	Synthesis	
Training	Predicted	Ground truth
Predicted	4.526 ± 0.066	4.449 ± 0.060
Ground truth	4.362 ± 0.066	4.522 ± 0.055

System	MOS
Tacotron 2 (Linear + G-L)	3.944 ± 0.091
Tacotron 2 (Linear + WaveNet)	4.510 ± 0.054
Tacotron 2 (Mel + WaveNet)	$\mathbf{4.526 \pm 0.066}$

- 정량적 평가 (MOS Score)
  - Tacotron 2: 4.53 ± 0.11
  - Human speech: 4.58 ± 0.08
  - Tacotron (Griffin-Lim): 3.82 ± 0.08

    → WaveNet conditioning으로 품질이 인간 수준에 근접.
- 정성적 평가
  - 발음 일관성, 억양, 자연스러움 모두 향상.
  - Attention alignment가 안정적으로 수렴.
  - 긴 문장에서도 자연스러운 prosody 유지.
- 속도
  - WaveNet inference는 여전히 실시간보다 느림.

6. Insight

- Tacotron 2는 텍스트 → 오디오 end-to-end TTS의 실현 가능성을 입증한 대표 연구.
- mel-spectrogram을 중간 표현으로 사용함으로써, Tacotron의 alignment 안정성과 WaveNet의 오디오 품질을 결합.
- 장점
  - 자연스러운 발음, 억양, 음색
  - 복잡한 파이프라인의 단순화
  - 인간 수준의 합성 품질 달성
- 한계
  - Autoregressive 구조로 인해 느린 inference 속도
  - 단일 화자 모델로 확장성 한정
- 후속 영향
  - Parallel WaveNet, WaveGlow, HiFi-GAN 등 non-autoregressive vocoder 개발의 촉매가 됨.
  - 이후 다화자·다언어 end-to-end TTS로 확장됨.