Tacotron : Towards End-to-End Speech Synthesis

https://arxiv.org/abs/1703.10135

0. Introduction

- 문자 입력만으로 음성을 직접 생성하는 end-to-end TTS 모델 제안함.
- 기존 TTS는 여러 모듈로 나뉘어 복잡하고 유지보수 어려움 있었음.
- Tacotron은 `<text, audio>` 쌍만으로 학습 가능해 전문 전처리 필요 없음.
- 기여점은 end-to-end 구조, 안정적 attention, 단일 모델 기반 고품질 음성 생성임.

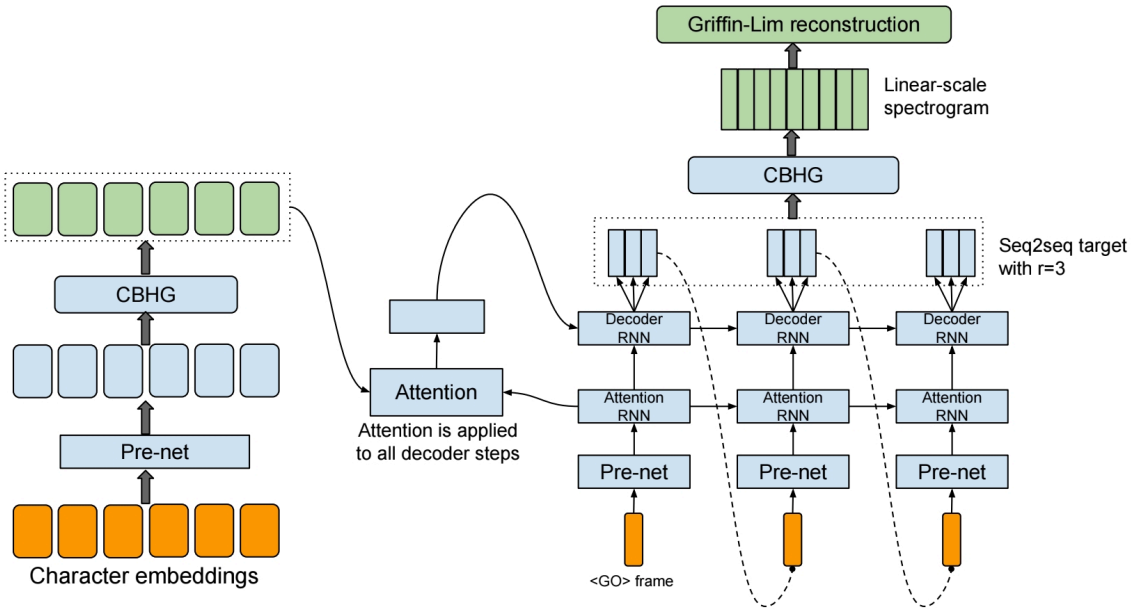
1. Overview

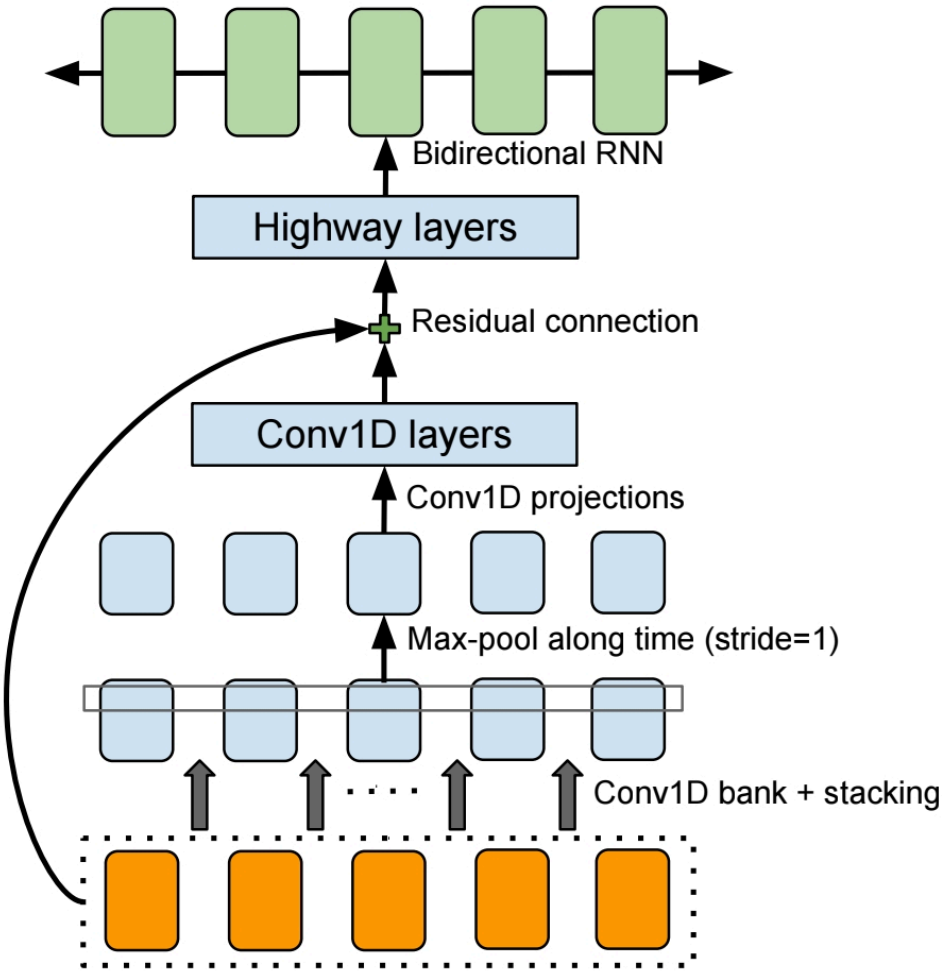
- seq2seq + attention 구조로 텍스트를 스펙트로그램으로 변환하는 방식임.
- 입력은 문자 임베딩, 출력은 멜/선형 스펙트로그램임.
- 인코더는 CBHG 모듈로 텍스트 특징 추출함.
- 디코더는 autoregressive로 프레임 단위 생성함.
- 최종 오디오는 Griffin–Lim으로 파형 복원함.
- 전체 파이프라인을 단일 모델로 처리하는 것이 핵심임.

2. Challenges

- 텍스트 길이와 오디오 길이가 다르므로 attention 불안정 문제 생길 수 있음.
- 같은 문장이라도 억양·속도 다양해 학습 난이도 높음.
- end-to-end 구조여도 품질–속도 균형 맞추기 어려움 있음.
- 데이터 품질·량에 민감하며 스타일 제어도 제한적임.

3. Method


Griffin-Lim reconstruction
Linear-scale spectrogram
CBHG
Seq2seq target with r=3
Decoder RNN
Decoder RNN
Decoder RNN
Attention
Attention RNN
Attention RNN
Attention RNN
Attention is applied to all decoder steps
Pre-net
Pre-net
Pre-net
<GO> frame
CBHG
Pre-net
Character embeddings



- 인코더 문자 임베딩 → CBHG → 텍스트 특징 생성함.
- 디코더 : attention 기반 RNN 구조 사용함.
- 프리넷으로 입력 안정화, 포스트넷으로 스펙트로그램 보정함.
- Griffin–Lim으로 최종 음성 재구성함.
- 추가 정렬 정보 없이 문자–음성 쌍만으로 학습함.

4. Experiments

- 단일 화자 영어 데이터셋 사용해 학습함.
- MOS로 음성 자연스러움 평가함.
- 기존 파라메트릭 TTS를 baseline으로 비교함.

- attention 안정성, 스펙트로그램 기반 생성 효율성도 검증함.

5. Results

Table 2: 5-scale mean opinion score evaluation.

	mean opinion score
Tacotron	3.82 ± 0.085
Parametric	3.69 ± 0.109
Concatenative	4.09 ± 0.119

- MOS 기준 3점대 후반 성능으로 기존 파라메트릭 TTS보다 우수함.
- 프레임 단위 생성 덕분에 샘플 단위 autoregressive 모델보다 빠름.
- CBHG, 프리넷, 포스트넷이 품질과 안정성에 기여함.
- Griffin–Lim 품질 한계 존재, 억양·표현력 제어는 제한적임.

6. Insight

- 복잡한 TTS 파이프라인을 단일 모델로 통합한 첫 주요 연구 중 하나임.
- 문자 기반 학습만으로도 자연스러운 음성 생성 가능함을 입증함.
- 단일 구조, 유지보수 용이, 빠른 생성, 높은 자연스러움.
- 보코더 한계, prosody 제어 부족, 데이터 요구량 큼.
- 고품질 vocoder 결합, 스타일·억양 제어, 다화자·다언어 확장, 데이터 효율 개선임.