Neural Speech Synthesis with Transformer Network

Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, Ming Liu

AAAI, 2019

임동주

목차

Introduction

- **Traditional TTS System**
  - Front end
    - Text analysis
    - Linguistic feature extraction
  - Back end
    - Speech synthesis

- **Tacotron1 / 2**
  - Text에서 바로 mel spectrograms 생성
  - 그 이후 음성 합성
    - Griffin Lim algorithm
    - WaveNet

- **Neural TTS**
  - RNN 계열 (Encoder + Decoder)
    - Encoder: input sequence를 semantic space에 매핑 후 hidden state 생성
    - Decoder: hidden state를 attention mechanism으로 문맥 정보로 인식
      decoder-hidden state 생성
      mel frame 반환

Introduction

- 기존 TTS 시스템의 문제점
  - 학습 효율 낮음
  - RNN 자체의 Long dependency 문제

- Neural TTS with Transformer의 장점
  - Self-attention을 반영
  - 마지막 hidden state의 sequential dependency를 해소
  - 전체적인 long distance dependency해결

Background

- Seq2Seq Model

- Tacotron2

- Transformer for NMT

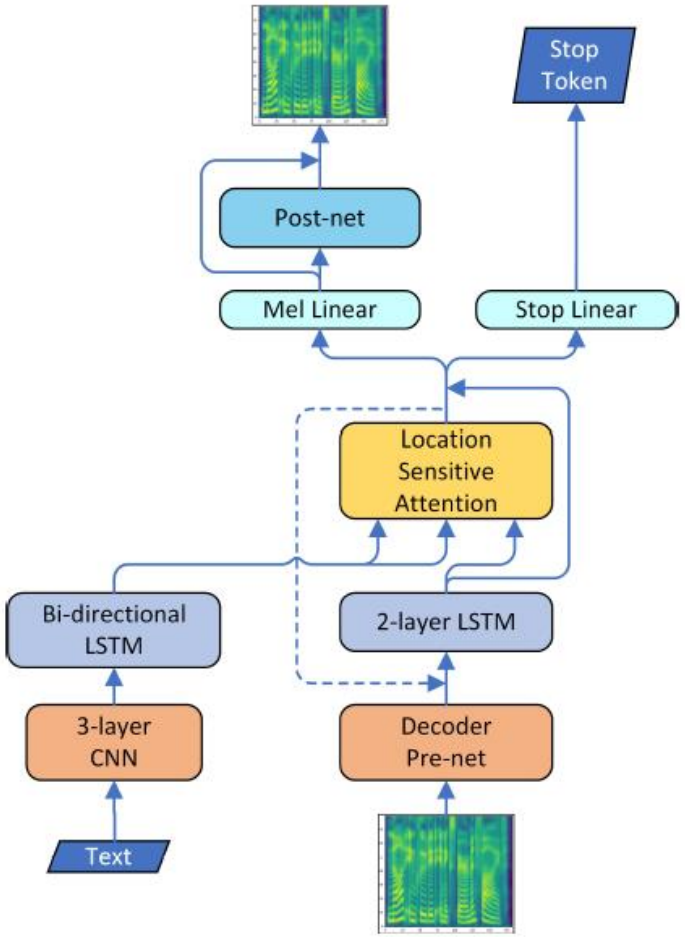
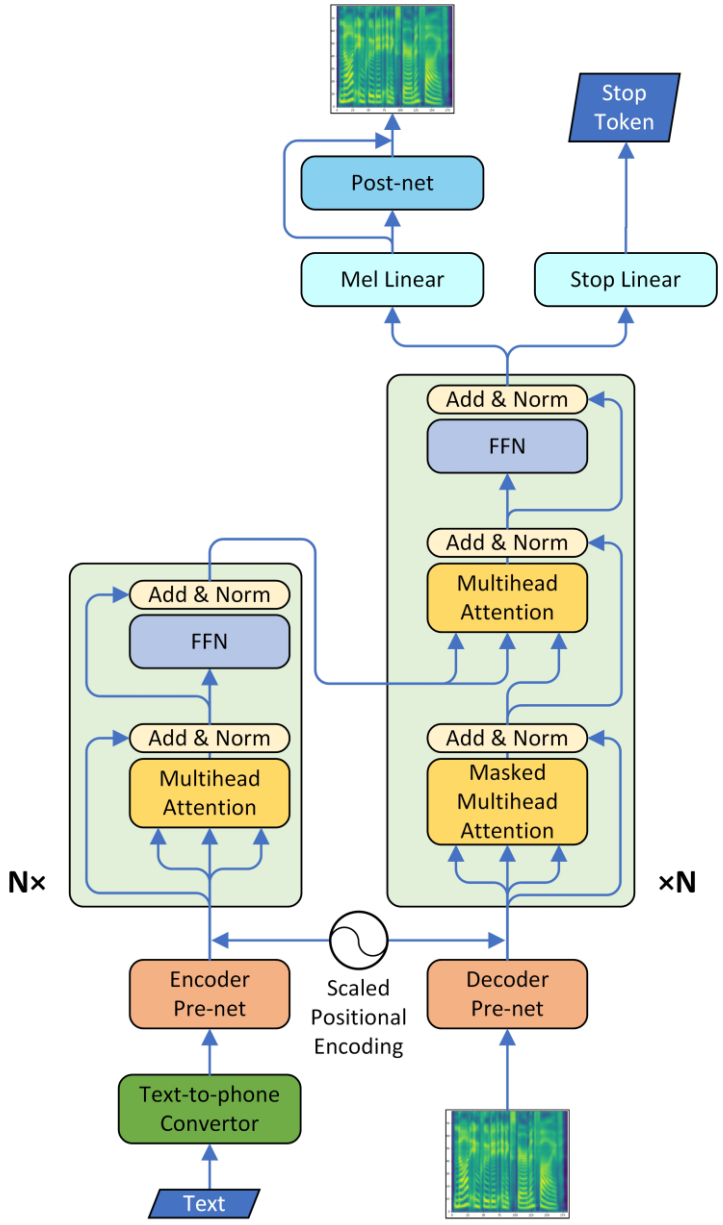


Figure 1: System architecture of Tacotron2.

Model Architecture



Method

1. Text-to-Phoneme Converter
    - Rule-Based Converter

2. Scaled Positional Encoding
    - 고정된 positional encoding을 사용할 경우, encoder / decoder pre-net 모두에게 과도한 부담을 줄 수 있음
    - 적절한 encoding의 반영을 위해 trainable weight를 적용

$$PE(pos, 2i) = \sin(\frac{pos}{10000^{\frac{2i}{d_{model}}}})$$
$$PE(pos, 2i+1) = \cos(\frac{pos}{10000^{\frac{2i}{d_{model}}}})$$

$$x_i = prenet(phoneme_i) + \alpha PE(i)$$

Method

3. Encoder Pre-net
    - 각 음소(Phoneme)는 512 임베딩 차원 보유
    - Tacotron2와 같은 3-layer CNN
    - Batch Normalization + ReLU

4. Decoder Pre-net
    - 2 FC Layer + ReLU에 Mel Spectrogram 입력 (256 hidden units)

        → Hidden size를 512로 설정할 경우, 성능 개선보다 수렴에 걸리는 비용이 더 큼

Method

5. Encoder
    - Transformer Encoder 사용
        - 전체를 한 번에 연산하기에 전체 시퀀스의 global context 반영 가능
          → audio prosody에 중요
        - Bi-directional RNN보다 학습 속도 향상

6. Decoder
    - Transformer Decoder 사용
        - 다양한 관점으로 encoder hidden states를 통합 가능
        - 더 나은 context vector 생성 가능
          → multi-head attention을 location sensitive 하게 변경을 시도했으나,
          학습 시간이 지나치게 많이 들고, 메모리 부족 현상이 쉽게 발생함

Method

7. Mel Linear, Stop Linear, Post-net
    - Tacotron2와 유사
    - 다만, stop linear projection에서 ‘stop’ positive 샘플의 부족으로 멈추지 않는 현상 발생
        - Positive 샘플에 5~8의 weight 반영하여 해결

Experiments

- Dataset
  - 25시간 여성 스피치 데이터(내부 데이터)

- Training Setup
  - Nvidia Tesla P100 * 4
  - Dynamic Batch Size(평균 16)

- Text-to-Phoneme Conversion and Preprocess
  - text-to-phoneme 변환을 통하여 pre-normalized phoneme 시퀀스를 입력으로 받음
    → Tacotron은 텍스트(character)를 그대로 입력으로 받음

- WaveNet Settings
  - 2 QRNN layers + 20 dilated layers

Experiments

- MOS & CMOS

- 랜덤으로 총 38개의 예제를 골라 20명의 네이티브 영어 스피커에게 MOS를 측정
  → 동점
- CMOS: Tacotron2 vs Transformer 둘 중 하나 선택
  → 본 논문의 모델이 살짝 우세

System	MOS	CMOS
Tacotron2	4.39 ± 0.05	0
Our Model	4.39 ± 0.05	**0.048**
Ground Truth	4.44 ± 0.05	-

Table 1: MOS comparison among our model, our Tacotron2 and recordings.

Experiments

- Mel Spectrogram 비교

- 논문에서 제안한 모델이 Ground Truth와 가장 유사
  - Layer 수도 중요한 것으로 보임

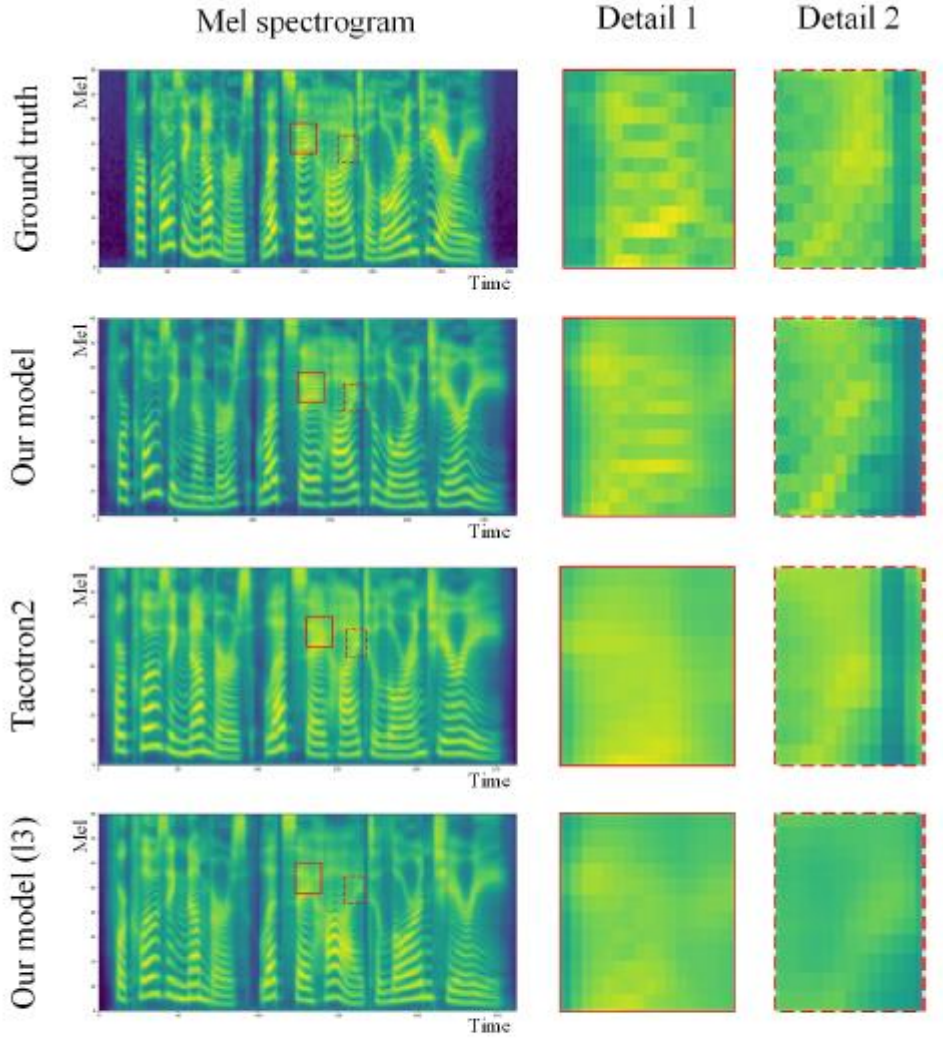


Figure 4: Mel spectrogram comparison. Our model (6-layer) does better in reconstructing details as marked in red rectangles, while Tacotron2 and our 3-layer model blur the texture especially in high frequency region. Best viewed in color.

Ablation Studies

- Re-centering Pre-net's Output
  - Re-centering Pre-net 적용한 것이 나은 성능 보임

Re-center	MOS
No	4.32 ± 0.05
Yes	**4.36** ± 0.05
Ground Truth	4.43 ± 0.05

Table 2: MOS comparison of whether re-centering pre-net's output.

- Different Positional Encoding Methods
  - Scaled PE 방식을 쓴 것이 나은 성능을 보임

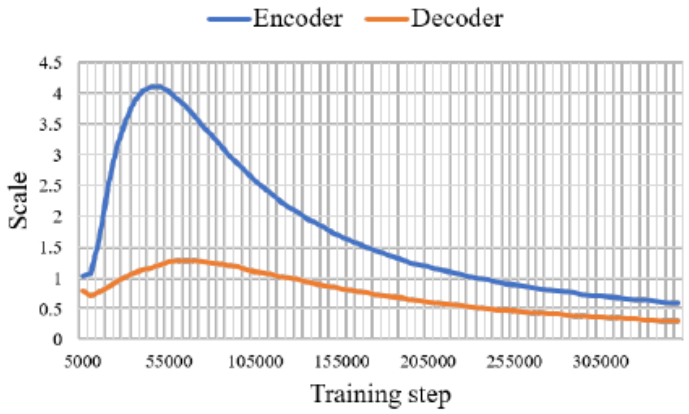


Figure 5: PE scale of encoder and decoder.

PE Type	MOS
Original	4.37 ± 0.05
Scaled	**4.40** ± 0.05
Ground Truth	4.41 ± 0.04

Table 3: MOS comparison of scaled and original PE.

Ablation Studies

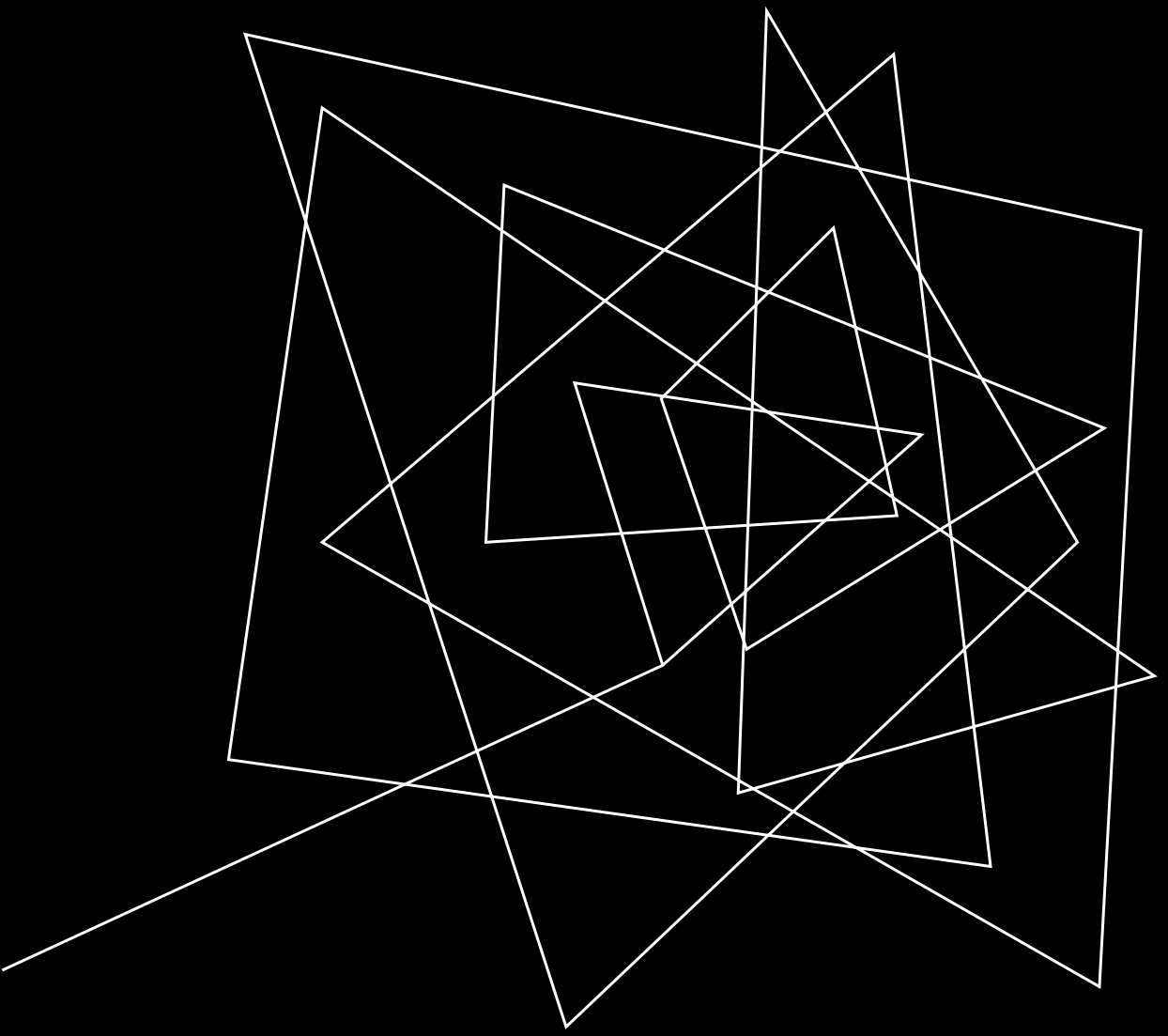
- Model with Different Hyper-Parameter
  - 논문에서 제안한 모델의 성능이 가장 좋음

Layer Number	MOS
3-layer	4.33 ± 0.06
6-layer	**4.41** ± 0.05
Ground Truth	4.44 ± 0.05

Table 4: Ablation studies in different layer numbers.

Head Number	MOS
4-head	4.39 ± 0.05
8-head	**4.44** ± 0.05
Ground Truth	4.47 ± 0.05

Table 5: Ablation studies in different head numbers.



Q&A