💚

1. NATURAL TTS SYNTHESIS BY CONDITIONING WAVENET ON MEL SPECTROGRAM PREDICTIONS

0. Abstract

- **Tacotron 2**에 대해 설명
  - 텍스트에서 직접 음성 합성을 수행하는 신경망 아키텍처
  - 문자 임베딩을 mel 스케일 스펙트로그램으로 매핑하는 순차적 시퀀스-시퀀스 특성 예측 네트워크로 구성
  - 이후 수정된 `WaveNet` 모델이 vocoder로 작용 → 시간 영역의 파형을 합성
- **실험 설계**
  - 설계 선택 사항을 검증하기 위해 시스템의 주요 구성 요소에 대한 제거 실험을 제시
  - 언어, 지속 시간 및 F0 특성을 활용하는 대신 `WaveNet`에 조건으로 mel 스펙트로그램의 영향을 평가
- 이 간결한 음향 중간 표현을 사용하면 `WaveNet` 아키텍처의 크기를 크게 축소할 수 있다는 것을 확인

▼ WaveNet

- 음성 합성과 음성 생성에 사용되는 신경망 모델 중 하나
- 특징
  1. 생성적 모델
     - WaveNet은 생성적 모델로 분류됨
       - 주어진 입력 정보를 기반으로 새로운 데이터를 생성하는 데 사용

        - 주로 음성 합성 및 음성 생성 작업을 위해 훈련되며, 입력으로 주어진 텍스트나 음성 신호에 따라 새로운 음성 신호를 생성할 수 있음

    2. 순환 신경망(RNN) 대신 합성곱 신경망(CNN)
        - WaveNet은 순환 신경망(RNN) 대신 합성곱 신경망(CNN)을 사용하여 시퀀스 데이터를 처리함

          ⇒ 긴 시퀀스에 대한 학습 및 생성이 더욱 효율적으로 이루어질 수 있음

    3. 고해상도 음성 생성
        - 고해상도 음성 신호를 생성할 수 있어서 자연스러운 음성 합성에 매우 적합함
        - 음성의 세부 사항과 풍부한 음향 특성을 재현할 수 있어서 인간과 거의 구별할 수 없는 음성을 생성할 수 있음

    4. 복잡한 확률 분포 모델링
        - 입력과 출력 간의 복잡한 확률 분포를 모델링하는 데 사용됨

          ⇒ 다양한 음성 특성을 생성하고 다양한 화자, 억양 및 언어 스타일을 흉내낼 수 있음

1. Introduction

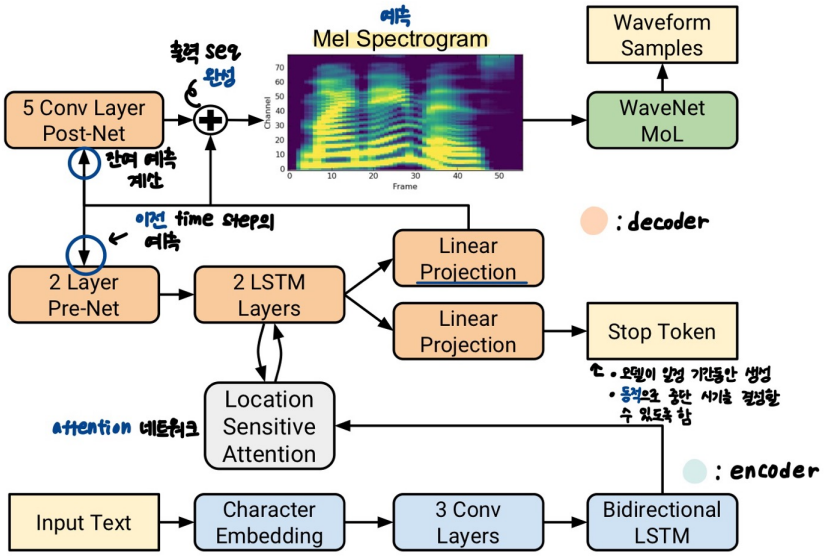
- 텍스트 음성 합성(TTS)은 여전히 어려운 과제임
    - 단위 선택을 통한 연결적 합성: 사전 녹음된 파형의 작은 단위를 조합하는 과정
    - 통계적 매개 변수 음성 합성: 음성 특성의 부드러운 궤도를 직접 생성하여 보코더에 의해 합성됨 → 연결적 합성이 경계 요소와 관련된 많은 문제를 해결
    - 시스템이 생성하는 오디오는 종종 인간의 음성과 비교하여 어두운 소리와 자연스럽지 않다고 들릴 수 있음
- WaveNet
    - 시간 영역의 파형을 생성하는 생성 모델 → 준수한 오디오 품질, 이미 일부 완전 TTS 시스템에서 사용되고 있음
    - 그러나 입력되는 정보(언어 특성, 예측된 로그 기본 주파수(F0), 및 음소 지속 시간)는 도메인 전문 지식을 필요로 함

    - 복잡한 텍스트 분석 시트템과 견고한 발음 가이드 등을 요구
- Tacotron
    - 문자 시퀀스에서 크기 스펙트로그램을 생성하기 위한 시퀀스-투-시퀀스 아키텍처 → 전통적인 음성 합성 과정을 단순화
    - 특정 artifact와 낮은 오디오 품질이 문제
- 해당 논문에서는 이러한 접근법들의 최상의 부분을 결합한 통합적이고 완전한 신경망 기반 음성 합성 방법을 제안
    - `Tacotron` 스타일 모델을 사용하여 mel 스펙트로그램을 생성하고 이를 수정한 `WaveNet` 보코더를 사용하여 자연스러운 음성을 합성
    - 논문의 방법은 높은 자연스러움과 품질을 제공하는 것으로 나타남

    ▼ mel 스펙트로그램

        - 음성 신호의 주파수 내용을 표현하는 시각화된 그래프나 데이터
        - 음성 처리 및 음성 분석 시 음성 특징 추출 및 합성에 사용
        - 주파수 스케일을 인간 청각 시스템의 특성에 더 가깝게 조정하기 위해 주파수 영역을 mel 스케일로 변환
        - 음성 신호의 주파수 내용을 시간에 따라 분석하는 데 사용 → 음성 특징 추출 및 음성 합성 모델에 입력으로 제공됨

2. 모델 구조



- 제안된 모델은 두 가지 구성 요소를 가지고 있음
  1. 입력 문자열 시퀀스로부터 mel 스펙트로그램 프레임의 시퀀스를 예측하는 어텐션 seq-to-seq 특성 예측 네트워크
  2. 예측된 mel 스펙트로그램 프레임에 의존하여 시간 영역 파형 샘플을 생성하는 WaveNet의 수정 버전

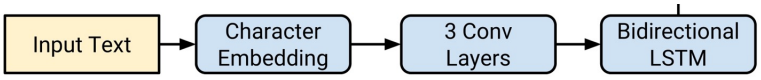
2-1. 중간 특성 표현

- 두 구성 요소 간의 연결을 위해 mel-주파수 스펙트로그램을 선택
  - 시간 영역의 파형에서 계산하기 쉬운 표현 → 두 구성 요소를 별도로 훈련할 수 있도록 함
  - 파형 샘플보다 부드럽고 각 프레임 내에서 위상에 대한 불변성을 가짐 → 제곱 오차 손실을 사용한 훈련이 용이
  - 선형 주파수 스펙트로그램과 관련이 있으며, 인간 청각 시스템의 응답을 모델링한 비선형 변환을 통해 얻어짐

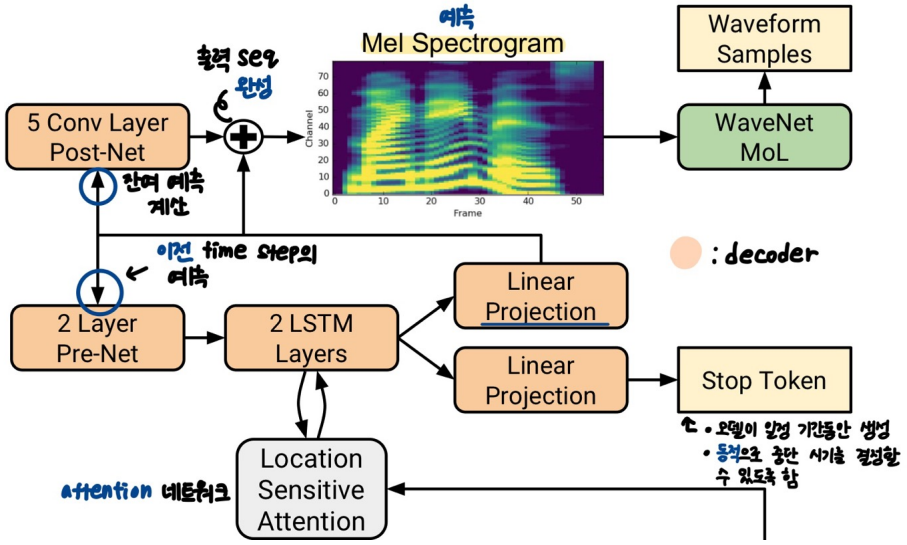
    ⇒ 주파수 내용을 더 적은 차원으로 요약하며, 특히 음성 가독성에 중요한 낮은 주파수 세부 사항을 강조하면서 고주파수 세부 사항을 덜 강조함
- mel 스펙트로그램은 일반적으로 위상 정보를 잃어버리기에 선형 스펙트로그램과는 달리 역 문제가 복잡하지만, `WaveNet`에서 사용되는 언어 및 음향 특성과 비교하면 오디오 신호의 더 단순한, 저수준 음향 표현임
  - mel 스펙트로그램을 조건으로 하는 유사한 `WaveNet` 모델을 사용하여 실질적으로 오디오를 생성하는 것이 가능
  - 실제로 수정된 WaveNet 아키텍처를 사용하여 mel 스펙트로그램에서 고품질 오디오를 생성할 수 있음을 확인할 수 있음

2-2. 스펙트로그램 예측 네트워크

- 해당 연구에서는 Tacotron과 유사한 구조를 가진 두 개의 주요 구성 요소로 구성됨
  1. 문자 시퀀스를 입력으로 받아 mel 스펙트로그램 프레임의 시퀀스를 예측하는 역할을 하는 seq-to-seq 특성 예측 네트워크(**Encoder**)
     - 입력 문자를 512차원의 문자 임베딩으로 표현하고, 이를 3개의 합성곱 레이어를 통과시키며 장기적인 컨텍스트를 모델링
     - 마지막 합성곱 레이어의 출력은 512개의 유닛을 가진 양방향 LSTM 레이어를 통과하여 인코딩된 특성을 생성



2. 인코딩 된 입력 시퀀스로부터 mel 스펙트로그램을 예측하는데 사용되는 수정된 `WaveNet` (**Decoder**)
    - 이전 시간 단계에서의 예측을 기반으로 현재 프레임의 mel 스펙트로그램을 하나씩 예측
    - 예측된 스펙트로그램은 최종적으로 5개의 합성곱 레이어로 구성된 Post-Net을 통해 전체 재구성을 향상시키기 위한 잔차로 추가됨



- 디코더의 출력과 어텐션 컨텍스트가 결합되어 정지 토큰을 예측하는 데 사용됨
    - 생성 시기를 동적으로 결정하여 항상 일정한 기간 동안 생성하지 않고 완료 시점을 동적으로 결정하기 위해 사용됨
- 네트워크의 일부 레이어에는 dropout 및 zone-out과 같은 정규화 기법이 적용됨
    - 출력 다양성을 도입하기 위해 드롭아웃은 추론 시간에만 특정 레이어에 적용됨
- **Tacotron 2**는 Tacotron과 유사한 구조를 가지고 있지만, 더 단순한 구성 요소를 사용하며 복잡성을 줄임
    - 단순한 LSTM 및 합성곱 레이어를 사용
    - CBHG 스택 및 GRU 순환 레이어를 대신 사용
    - 축소 요인을 사용 x
    - 각 디코더 단계가 단일 스펙트로그램 프레임에 해당

2-3. WaveNet Vocoder

- **Tacotron2**는 `WaveNet` 아키텍처를 수정하여 mel 스펙트로그램 특성을 시간 영역의 파형 샘플로 변환함
  - 수정된 아키텍처는 30개의 확장된 컨볼루션 레이어로 구성되어 있으며, 이를 3개의 확장 주기로 그룹화
    → 각 레이어의 확장 비율은 2k (mod 10)로 설정됨
  - 스펙트로그램 프레임의 12.5ms 프레임 간격을 처리하기 위해 조건 스택에는 3개 대신 2개의 upsampling 레이어만 사용됨
- 확률 버킷을 예측하는 대신 `PixelCNN++` 및 `Parallel WaveNet`을 따라 10개 구성 요소의 로지스틱 분포 혼합(MoL)을 사용하여 16비트 샘플을 24kHz에서 생성

  ▼ 로지스틱 분포 혼합(MoL)

  - 확률 분포 모델링 기법 ⇒ 데이터의 분포를 여러 개의 로지스틱 분포로 구성된 혼합 분포로 표현
  - 특징

    1. 혼합 분포
       - 여러 로지스틱 분포를 합쳐서 하나의 분포로 표현
       - 각 로지스틱 분포는 서로 다른 모수(평균, 스케일)를 가지며, 이러한 분포들의 혼합을 통해 다양한 데이터 패턴을 모델링할 수 있음
    2. 유연성
       - 다양한 데이터 분포를 모델링할 수 있는 유연성을 제공
       - 데이터가 여러 모드(다중 군집)를 가지거나 복잡한 분포를 따를 때 유용하게 적용됨
    3. 매개 변수 예측
       - 각 로지스틱 분포의 매개 변수(평균, 스케일)를 예측하는데 사용됨

         ⇒ 모델은 주어진 입력 데이터에 대한 가장 적합한 로지스틱 분포를 선택하고 해당 분포로부터 샘플을 생성할 수 있음

  - 로지스틱 혼합 분포를 계산하기 위해 WaveNet 스택 출력은 `ReLU` 활성화를 거친 후 각 혼합 구성 요소의 매개 변수(평균, 로그 스케일, 혼합 가중치)를 예측하기 위한 선형 투영(linear projection)을 거침
  - 손실: 실제 샘플의 음의 로그 우도로 계산됨

3. 실험 & 결과

3-1. 훈련 setup

- 먼저 특성 예측 네트워크를 독립적으로 훈련한 다음, 첫 번째 네트워크에서 생성된 출력을 사용하여 수정된 `WaveNet`을 독립적으로 훈련하는 과정으로 구성됨
- 특성 예측 네트워크를 훈련하기 위해 표준 최대 우도 훈련 절차를 적용
  - 디코더 측의 예측된 출력 대신 올바른 출력을 공급하는 방식
    ⇒ 교사 강제라고도 함
  - 단일 GPU에서 `batch_size = 64`로 수행
  - 옵티마이저: `Adam`
    - 0.9와 0.999로 시작하는 γ 값
    - 학습률: 10^{-6}
  - 학습률: 10^{-3}(50,000 번의 반복 이후부터 지수적으로 감소)
  - 10^{-6}의 L2 정규화 적용
- 이후 수정된 `WaveNet`을 훈련
  - 특성 예측 네트워크의 예측된 출력 대신 올바르게 정렬된 예측을 사용
    - 예측 네트워크가 교사 강제 모드에서 실행되며, 각 예측된 프레임이 인코딩된 입력 시퀀스와 대상 스펙트로그램의 이전 프레임에 조건을 걸게 됨

      ⇒ 각 예측된 프레임이 대상 파형 샘플과 정확하게 일치하게 됨
  - 32개의 GPU에 분산된 `batch_size = 128`로 수행
    - 동기화된 업데이트를 사용
  - 옵티마이저: `Adam`
    - γ: 0.9, 0.999
    - 학습률: 10^{-4}(고정)
  - 최근 업데이트의 모델 가중치를 평균화하는 것이 품질 향상에 도움이 됨
    - 네트워크 매개 변수의 지수 가중 이동 평균을 유지 → 추론에 사용
    - 수렴 속도를 높이기 위해 파형 대상을 초기 출력과 미래 분포에 더 가깝게 만들기 위해 127.5 배율로 조정
- 모든 모델을 내부 US 영어 데이터 세트에서 훈련

- 단일 전문 여성 화자의 24.6시간의 음성이 포함되어 있음
- 모든 텍스트는 철자로 표시

⇒ 정규화된 텍스트로 훈련

3-2. 평가

- 추론 모드에서 음성을 생성할 때, 훈련 중에 사용한 교사 강제 구성과는 달리 이전 단계의 예측된 출력을 디코딩 중에 입력으로 사용
- 평가를 위해 테스트 세트에서 100개의 예제를 선택하고 이를 인간 평가 서비스에서 평가
  - 평가자들은 1부터 5까지의 범위에서 각 샘플을 평가 ⇒ 주관적 평균 의견 점수(MOS)를 계산
  - 이를 통해 음성 합성 시스템의 성능을 평가
- 모든 모델은 동일한 데이터로 훈련되었으며, 여러 이전 시스템과 비교하여 제안된 시스템이 다른 TTS 시스템을 능가하며 실제 음성과 유사한 MOS를 달성

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 실제 오디오	4.582 ± 0.053
Tacotron 2 (this paper)	$\mathbf{4.526 \pm 0.066}$

Table 1. Mean Opinion Score (MOS) evaluations with 95% confidence intervals computed from the t-distribution for various systems.

- side-by-side 평가에서는 우리 시스템으로 생성된 오디오와 실제 음성 간의 선호도를 비교
  - 실제 데이터에 대한 선호도가 약간 더 높음 ⇒ 가끔의 발음 오류가 주된 이유로 지적됨

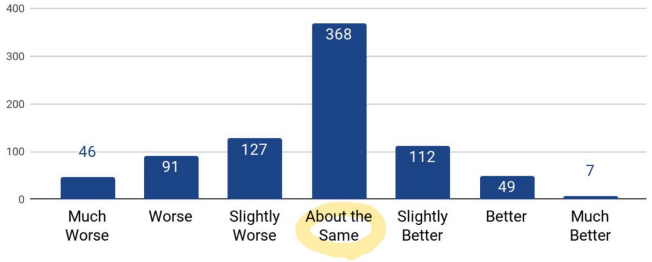


Fig. 2. Synthesized vs. ground truth: 800 ratings on 100 items.

- 테스트 세트에서 Tacotron 2는 MOS 4.354를 얻어냄
  - 에러 분석에서 반복 단어를 포함한 문장이 없음
  - 발음 오류, 건너 뛴 단어, 부자연스러운 억양을 포함한 에러가 있었음

    ⇒ 억양 모델링 측면에서 개선이 필요함
- 도메인 밖(뉴스 제목 등) 텍스트에 대한 37개 샘플에서 Tacotron 2는 MOS 4.148를 얻음
  - 언어적 특성에 따라 조건을 걸은 WaveNet은 MOS 4.137를 얻었음

⇒ Tacotron 2는 더 자연스러우나 발음 상의 어려움이 종종 발생함을 보여주며, end-to-end 접근 방식의 한계를 나타냄

3-3. 실험적 연구

3-3-1. 예측된 특성 vs 실제값

- 우리 모델의 두 구성 요소(= encoder, decoder)는 별도로 훈련되었지만, WaveNet 구성 요소는 훈련을 위해 예측된 특징에 의존
  - 대안) mel 스펙트로그램에서 독립적으로 WaveNet을 훈련
- **실험 결과**

	Synthesis	
Training	Predicted	Ground truth
$\hat{Y}$ Predicted	4.526 ± 0.066	4.449 ± 0.060
Y Ground truth	4.362 ± 0.066	4.522 ± 0.055

  - 예상대로, 훈련에 사용된 특징이 추론에 사용된 특징과 일치할 때 최상의 성능을 얻음
  - 그러나 예측된 특징에서 합성하도록 훈련된 모델은 반대 경우보다 성능이 떨어짐

    ⇒ 예측된 스펙트로그램이 실제값보다 과도하게 스무딩되고 덜 자세하다는 경향 때문
  - 이는 특성 예측 네트워크에서 최적화된 제곱 오차 손실의 결과
    - 실제값 스펙트로그램에서 훈련된 경우, 네트워크는 과도하게 스무딩된 특징에서 고품질 음성 파형을 생성하는 학습을 하지 않음

3-3-2. 선형 스펙트로그램

- mel 스펙트로그램 예측 대신 선형 주파수 스펙트로그램 예측을 실험
- Griffin-Lim을 사용하여 스펙트로그램을 역전시킬 수 있도록 함
    - `WaveNet`은 Griffin-Lim과 비교하여 훨씬 높은 음질의 오디오를 생성하지만, 선형 스케일 또는 mel 스케일 스펙트로그램의 사용 사이에 큰 차이가 없음

      ⇒ mel 스펙트로그램의 사용은 더 간결한 표현이기 때문에 엄격히 더 나은 선택으로 보임

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

성능 차이가 거의 x

3-3-3. Post-Processing 네트워크

- 미래 프레임의 정보를 디코딩하기 전에 사용할 수는 없음
    - 과거 및 미래 프레임을 포함시키기 위한 컨볼루션 후처리 네트워크를 사용하여 특징 예측을 개선
- `WaveNet` 자체가 이미 컨볼루션 레이어를 포함하고 있음

  ⇒ `WaveNet`을 vocoder로 사용할 때에도 여전히 후처리 네트워크가 필요한지를 실험
    - 후처리 네트워크가 있는 경우와 없는 경우의 모델을 비교
        - 후처리 네트워크 없이는 모델이 4.429 ± 0.071의 MOS 점수를 얻음
        - 후처리 네트워크가 있는 경우 4.526 ± 0.066의 MOS 점수를 얻음

  ⇒ 후처리 네트워크가 네트워크 설계의 중요한 부분임을 의미

3-3-4. WaveNet 단순화 시키기

- `WaveNet`의 특징 중 하나는 수렴을 기하급수적으로 증가시키기 위해 확장된 컨볼루션(dilated convolution)을 사용하는 것임
    - 수용 영역 크기와 레이어 수가 다른 모델을 평가

      ⇒ mel 스펙트로그램이 이미 프레임 간의 장기 종속성을 캡처하고 있으므로 작은 수용 영역과 얕은 네트워크가 문제를 충분히 해결할 수 있을 것이라는 가설을 테스트
- **실험 결과**

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

Table 4. WaveNet with various layer and receptive field sizes.

- 12개 레이어와 10.5 ms 수용 영역으로도 고품질 오디오를 생성할 수 있으며, 기준 모델에서는 30개 레이어와 256 ms를 필요로 함
- 오디오 품질에 큰 수용 영역 크기가 필수적인 요소가 아니라는 점을 확인할 수 있음
  - 복잡성 감소를 허용하는 것은 mel 스펙트로그램에 조건을 거는 선택이라는 가설을 제기
- 반면 확장된 컨볼루션을 완전히 제거하면 수용 영역이 기준과 비교하여 두 개의 크기 순서 작아지고 품질이 크게 저하됨
  - 모델이 웨이브폼 샘플의 시간 스케일에서 충분한 문맥을 필요로 하며 고품질 사운드를 생성하기 위함을 의미

4. 결론

- **Tacotron 2**를 설명
  - seq-to-seq 반복 네트워크와 어텐션을 결합하여 수정된 `WaveNet` 보코더와 함께 mel 스펙트로그램을 예측하는 완전히 신경망 기반의 텍스트 음성 변환 (TTS) 시스템
  - `Tacotron` 수준의 억양과 `WaveNet` 수준의 오디오 품질로 음성을 합성
  - 복잡한 특성 엔지니어링에 의존하지 않고 데이터로 직접 훈련할 수 있음
  - 자연스러운 인간 음성과 유사한 최신 음향 품질을 달성