논문 21-46-02-03 The Journal of Korean Institute of Communications and Information Sciences '21-02 Vol.46 No.02
https://doi.org/10.7840/kics.2021.46.2.217

리포머 네트워크를 이용한 종단형 한국어 음성 합성 시스템

임 형 래•, 천 성 준*, 최 병 진*, 김 민 찬*, 김 남 수°

End-to-End Korean Speech Synthesis System Using Reformer Network

Hyeong Rae Ihm•, Sung Jun Cheon*, Byoung Jin Choi*, Min Chan Kim*, Nam Soo Kim°

요 약

본 논문에서는 리포머 네트워크(Reformer network)를 사용하여 설계한 종단형 한국어 음성 합성 시스템을 제안한다. 종단형 음성합성 모델 중 높은 성능을 보이는 트랜스포머 네트워크의 메모리 비효율을 해결하기 위해 리포머 네트워크로 대체하여 사용하였고, 스펙트로그램과 한국어 텍스트 시퀀스 길이 간의 불균형이 어텐션 에너지 추정에 주는 부정적인 영향을 텍스트 시퀀스 길이 확장을 통해 해결하였다. 텍스트 시퀀스 샘플을 반복하여 길이를 확장하는 방법과 샘플 사이의 연결 정보를 추가하여 길이를 확장하는 방법을 사용하였다. 실험 결과 리포머 네트워크를 사용한 음성 합성 시스템이 트랜스포머 네트워크 기반 음성합성 시스템에 비해 적은 메모리로 학습이 가능하며 텍스트 샘플 간의 연결 정보를 사용하여 자연스러운 음성을 생성할 수 있음을 확인하였다.

Key Words : end-to-end speech synthesis, reformer network, text expansion communication, signal processing, Neutral systems, Communication Sciences, Network

ABSTRACT

In this paper, we propose a End-to-end Korean speech synthesis system using a reformer network. Transformer TTS shows high performance among the end-to-end speech synthesis models, but has the memory inefficiency in the training stage. In order to solve the memory inefficiency, Transformer network was replaced with the Reformer network. In addition, the imbalance between the spectrogram and the length of the Korean text sequence had a negative effect on the attention energy estimation. A method of extending the length by repeating text sequence samples and a method of extending the length by adding connection information between samples were used. As a result of the experiment, it was confirmed that the speech synthesis system using the reformer network can be trained with relatively less memory and can generate natural speech using connection information between text samples.

※ 이 연구는 방위사업청 및 국방과학연구소의 재원에 의해 설립된 신호정보 특화연구센터 사업의 지원을 받아 수행되었음.

• First Author : Seoul National University Department of Electrical and Computer Engineering and Institute of New Media and Communications and Human Interface Laboratory, hrim@hi.snu.ac.kr(석사), 학생회원

° Corresponding Author : Seoul National University Department of Electrical and Computer Engineering and Institute of New Media and Communications and Human Interface Laboratory, nkim@snu.ac.kr, 종신회원

* Seoul National University Department of Electrical and Computer Engineering and Institute of New Media and Communications and Human Interface Laboratory, 학생회원

논문번호 : 202009-222-A-RU, Received September 10, 2020; Revised October 7, 2020; Accepted October 8, 2020

Ⅰ. 서 론

최근 딥 러닝 기술을 바탕으로 하여 음성 신호처리 분야의 연구가 활발히 진행되었고 관련된 여러 분야에서 성능이 크게 개선되고 있다. 음성합성 분야에서는 딥 러닝 기술이 사용되기 이전에 제안되었던 고전적인 음성합성 모델을 보완하여 성능이 크게 개선된 모델이 제안되고 있다. 고전적인 음성합성 시스템은 텍스트 시퀀스를 문자 단위 혹은 음소 단위의 소단위 시퀀스로 변환한 후 발음 특성을 추출하고, 각 텍스트의 발음 길이를 추정하는 기간(duration) 모델과 음성 신호의 특징을 추정하는 음향 모델을 통해 추정한다. 여기서 기간 모델과 음향 모델은 HMM(Hidden Markov Model) 기반의 파라메트릭 통계 모델을 사용한다.[1, 2] 추정된 음성 특성 시퀀스는 보코더(vocoder)를 통해 음성 샘플 시퀀스로 변환된다. 딥 러닝 프레임워크를 사용하여 제안된 모델은 고전적인 음성합성 모델의 일부 모듈을 대체하는 방식이 제안된 바 있다.

이후에는 텍스트 시퀀스로부터 음성 특성 시퀀스를 추정하는 종단형 음성합성 시스템이 제안되었다. 딥 보이스(Deep Voice)[3, 4], 타코트론(Tacotron)[5, 6], 트랜스포머 TTS(Transformer TTS)[7] 등의 모델이 제안되었고, 기존의 고전 음성합성 모델이 복잡한 설계를 필요로 하며, 다단계의 구조로 인해 오차가 누적되는 문제를 해결하였다. 특히, 타코트론과 트랜스포머 TTS 모델은 어텐션(attention) 기반의 구조를 사용하여 텍스트와 음향 시퀀스 샘플들을 대응시켰는데, 이를 이용하면 별도의 발음 구간 정보를 주지 않더라도 학습 과정에서 두 시퀀스의 대응을 수행할 수 있다는 장점이 있다. 또한, 상호 정보량을 사용하여 어텐션 기반 스타일 음성 합성이 가능한 모델이 제안된 바 있다[8].

어텐션 기반의 음성합성 모델은 매 쿼리 시퀀스마다 키 시퀀스 전체에 대해 어텐션 에너지 값을 계산하게 된다. 그 결과, 메모리 복잡도는 쿼리 시퀀스 길이를 L_q, 키 시퀀스 길이를 L_k라고 했을 때, $O(L_qL_k)$의 값을 갖게 되어 메모리 측면에서 비효율성을 갖는다는 문제가 있다. 특히 트랜스포머 TTS에는 인코더의 재귀-어텐션(self-attention), 디코더의 재귀-어텐션과 인코더-디코더 어텐션이 각각 3번씩 중첩된 형태를 갖고 있어 학습 시 큰 규모의 메모리 사용을 필요로 한다. 또한 인코더와 디코더가 중첩된 형태에서 역전파 과정에서 활성화 값을 필요로 하는데, 매 중첩마다 활성화 값을 저장해야 하므로 학습과정에서의 메모리 비효율을 악화시킨다. 한편, 트랜스포머 TTS에서 미니 배치 사이즈는 안정적인 학습에 중요한 역할을 한다고 보고된 바 있다. 따라서 메모리를 효율적으로 사용하여 어텐션 기반 모델을 학습할 수 있다면 더 안정적인 학습이 가능하다.

최근 제안된 리포머 네트워크[9]는 위치 민감성 해싱[10] 어텐션(locality-sensitive hashing attention)을 사용하여 어텐션 연산 과정에서 메모리 측면에서 효율적인 계산을 수행한다. 또한, 가역 잔여 네트워크(reversible residual network)[11]를 사용하여 잔여 네트워크[12]의 중첩 시 신경망의 각 레이어의 활성화 값을 저장해야하는 단점을 해소하여 메모리 측면에서의 효율성을 강화하였다. 두 방식을 사용하여 구성한 리포머 네트워크는 자연어 이해 분야에서 성능의 큰 저하 없이 메모리 사용을 낮춰 효율적인 학습이 가능함이 보고되었다.

본 논문에서는 리포머 네트워크를 이용한 한국어 음성합성 시스템을 제안한다. 디코더의 재귀-어텐션에 위치 민감성 해싱 어텐션과 가역 잔여 네트워크를 사용하여 효율적으로 메모리를 사용하여 학습을 수행하였고, 전진 어텐션(forward attention)[13]을 사용하여 위치 민감성 해싱 어텐션이 시퀀스 전체에 대해 에너지 값을 계산하지 않아 텍스트와 스펙트로그램간의 정렬이 불안정할 수 있는 점을 보완하였고, 인코더-디코더 어텐션이 단조증가 형태를 보다 쉽게 형성할 수 있도록 하였다. 또한, 한국어 텍스트 처리 과정에서 초성, 중성, 종성 각각의 단일 자소만을 사용하지 않고, 직후의 자소에 대한 정보를 기존 시퀀스 샘플 사이에 추가함으로써 보다 자연스러운 발음을 위한 텍스트 처리를 수행하였다. 실험은 한국어를 발음하는 약 10시간 분량의 남성 단일 화자 데이터베이스를 사용하였으며, 트랜스포머 TTS 모델과, 초, 중, 종성 자소 시퀀스를 이용한 한국어 처리 모듈을 사용한 리포머 TTS를 비교 모델로 설정하여 본 논문에서 제안한 방식의 타당성을 검증하였다.

Ⅱ. 본 론

2.1 리포머 네트워크 기반 음성합성 시스템

본 항에서는 리포머 네트워크에 대한 설명과 리포머 네트워크를 기반으로 한 음성합성 시스템에 대해 설명한다.

2.1.1 리포머 네트워크

트랜스포머 네트워크는 어텐션 기반 시퀀스-투-시

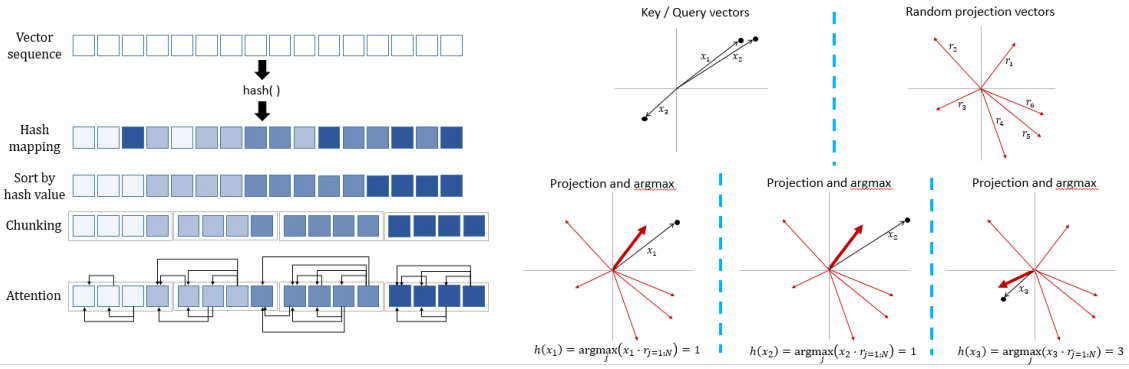

그림 1. 위치-민감성 해싱 어텐션
Fig. 1. Locality-sensitive hashing attention

퀀스 딥 러닝 모델로 어텐션은 두 시퀀스 샘플 간의 상관관계를 쿼리와 키 시퀀스의 모든 샘플에 대해서 계산하게 된다. 따라서 인코더와 디코더의 어텐션마다 쿼리의 시퀀스 길이를 L_q, 키의 시퀀스 길이를 L_k라고 했을 때, 메모리 복잡도는 $O(L_qL_k)$의 값을 갖게 된다.

하지만, 학습된 트랜스포머 모델의 어텐션은 전체 시퀀스 중 일부를 제외하면 어텐션 가중치 값이 0에 가깝게 나타난다. 시퀀스 중 가중치 값이 0에 가까이 나타나는 샘플들에 대해 어텐션 계산을 하지 않으면 비효율적인 메모리 소모를 줄일 수 있다. 위치 민감성 해싱 어텐션을 사용하게 되면 시퀀스 샘플들 중 가중치 값이 0으로 예상되는 샘플들을 제외하고 어텐션을 수행함으로써 메모리 소모를 절약하게 된다. 또한, 인공 신경망의 학습 시 역전파 과정에서 경사도를 계산할 때 잔여 네트워크가 존재할 경우 해당 잔여 레이어의 활성화 값을 저장해 두어야 하는데, 트랜스포머 모델은 잔여 네트워크가 중첩된 구조이기 때문에 매 레이어에서 잔여 레이어의 활성화 값을 역전파 과정에서 메모리에 별도로 저장하여야 한다. 잔여 레이어의 활성화 값을 저장하지 않고 직접 구할 수 있는 가역 잔여 네트워크를 사용하면 메모리를 절약하여 신경망 학습을 진행할 수 있다.

(1) Locality-sensitive hashing attention

리포머 네트워크는 위치 민감성 해싱 어텐션을 사용하여 어텐션 가중치가 높다고 추정되는 키 벡터들 사이의 군집화를 통해 비효율적인 메모리 소모를 막을 수 있다. 그림 1과 같이 서로 정보가 비슷한 벡터가 같은 해시 함수 출력을 갖는 해시 알고리즘을 사용하여 쿼리와 키의 정보가 비슷한 것들이 같은 해시 값을 갖게 하여 이에 따라 군집화 함으로써 같은 군집 내에서 어텐션을 계산하도록 한다. 같은 해시 값을 갖는 군집의 쿼리, 키 벡터들 안에서 어텐션 계산을 수행한다.

(2) Reversible residual network

잔여 네트워크[12]는 딥 러닝 프레임워크 기반 모델의 경사 소실(vanishing gradient) 문제를 해소하는 방법으로 제안되었던 네트워크이며 잔여 네트워크를 기반으로 설계된 여러 모델이 인공 신경망 네트워크의 여러 분야에서 우수한 성능을 보인 바 있다. 하지만, 잔여 네트워크는 학습 시 역전파 과정에서 잔여 네트워크 출발 지점의 신경망의 활성화 값을 저장해두어야 경사 계산이 가능하다. 이후 제안된 가역 잔여 네트워크[11]는 잔여 네트워크의 특성을 유지하면서 역전파 계산 시 해당 위치의 신경망 활성화 값을 순차적으로 계산해 나가면서 메모리에 저장해두지 않고 계산이 가능하다.

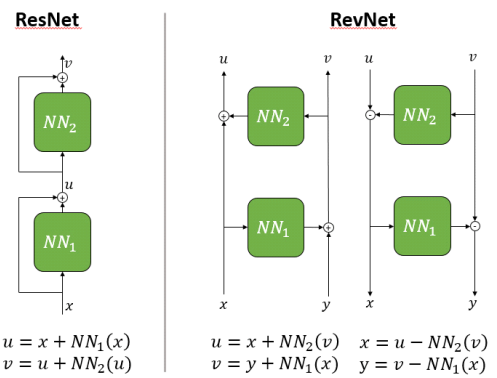

그림 2. 가역 잔여 네트워크
Fig. 2. Reversible residual network

그림 2는 잔여 네트워크와 가역 잔여 네트워크의 구조이다. 그림 2의 오른쪽 구조를 사용하면 역전파 과정에서 잔여 네트워크 입력단의 신경망 활성화 값을 잔여 네트워크의 합 지점의 신경망 활성화 값을 이용하여 구할 수 있다. 이를 통해 블럭이 중첩된 구조인 인코더와 디코더에서 반복적으로 구성되는 잔여 네트워크에서 역전파시 매 잔여 네트워크마다 신경망 활성화 값이 저장되는 비효율성을 해소할 수 있다.

2.1.2 리포머 네트워크 기반 음성합성 모델 구조

위치 민감성 해싱 어텐션과 가역 잔여 네트워크를 사용한 리포머 네트워크를 이용하여 음성합성 모델을 구성하여 학습 과정에서 메모리를 효율적으로 사용할 수 있다. 트랜스포머 TTS에서 미니 배치 사이즈가 안정적인 학습에 핵심적인 요소로 언급된 바 있으며, 리포머 네트워크를 사용하여 메모리를 효율적으로 사용한다면 충분한 미니 배치 사이즈를 확보할 수 있을 것이다. 본 논문에서 제안하는 구조는 그림 3과 같다.

인코더에서는 텍스트 전처리 및 샘플 단위의 확장을 통해 얻은 텍스트 정보 시퀀스를 입력으로 받아서 256차의 멀티-헤드 재귀 어텐션과 잔여 네트워크를 통과하게 된다. 이러한 block을 3번 중첩하여 모델의 표현력을 높였으며, 인코더의 최종 출력을 디코더의 인코더-디코더 어텐션의 입력으로 사용하게 된다.

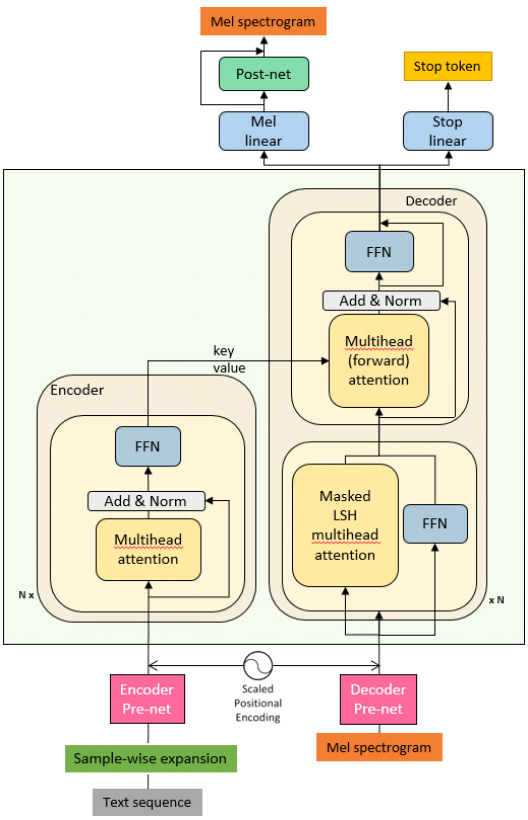

그림 3. 모델 전체 구조
Fig. 3. Model architecture

디코더에서는 정답 음성 신호로부터 추출된 멜 스펙트로그램과 인코더를 거친 텍스트의 컨텍스트 시퀀스를 이용하여 다음 시간 스텝의 멜 스펙트로그램을 추정하게 된다. 군집의 크기 32, 256차의 위치 민감성 해싱 어텐션이 트랜스포머 네트워크 디코더의 재귀 어텐션을 대체하며, 가역 잔여 네트워크를 거쳐 인코더-디코더 어텐션의 입력으로 사용된다. 위치 민감성 해싱 어텐션은 인코더-디코더 어텐션에서는 멜 스펙트로그램과 텍스트의 의미차로 인해 사용하지 않았다. 위치 민감성 해싱 어텐션은 쿼리 시퀀스와 키 시퀀스의 거리가 가까운 점을 이용하게 되는데, 재귀 어텐션과 달리 쿼리 시퀀스와 키 시퀀스가 서로 다른 종류의 정보를 갖고 있기 때문에, 벡터 상에서 거리가 가깝다고 하더라도 두 벡터는 연관이 없을 가능성이 크기 때문이다. 인코더-디코더에서는 256차의 멀티-헤드 어텐션을 사용하였고, 그 후 선형 신경망과 잔여 네트워크를 통해 멜 스펙트로그램과 스탑 토큰을 추정하는 출력을 내보내게 된다.

한편, 위치 민감성 해싱 어텐션은 query 샘플에 대해 key 시퀀스의 일부에 대해 어텐션 계산을 수행하기 때문에 트랜스포머 네트워크의 재귀-어텐션에 비해 다소 불안정할 수 있다. 한편, 음성합성의 특성상 모든 텍스트 시퀀스에 대해 순차적으로 반복이나 생략 없이 단조증가 형태의 어텐션을 형성해야 하는데, 이를 보완하기 위해 어텐션의 단조증가 특성이 쉽게 나타나도록 하는 전진 어텐션 알고리즘을 사용한다.

디코더에서 생성된 출력은 다음 단계의 멜 스펙트로그램과 스탑 토큰을 추정하는 신경망의 입력으로 사용된다. 멜 스펙트로그램은 256차의 선형 신경망과 5의 커널 크기를 갖는 5개의 256차 컨벌루션 네트워크를 사용하고 선형 신경망의 출력과의 잔여 네트워크를 거쳐 추정된다. 정답 멜 스펙트로그램과의 L1 loss를 사용하며, 스탑 토큰은 256차의 선형 신경망을 거쳐 발화의 종점인지 여부를 결정하는 1차 출력을 추정하며 이진 크로스 엔트로피 손실을 사용하여 학습을 진행하게 된다. 다음 스텝의 멜 스펙트로그램을 계산할 때는 현재 스텝까지의 멜 스펙트로그램을 사용하여 시간에 대해 인과적으로 계산하며, 스탑 토큰은 실제 음성 생성 시 스탑 토큰을 매 스텝에서 계산하여 일정 확률 이상 값으로 나타날 경우 샘플 생성을

중단하고 스펙트로그램 추정을 마치게 된다.

2.2 한국어 음성합성을 위한 텍스트 전처리

음성합성을 위해 한글 텍스트를 입력으로 받을 경우 한글 음절은 여러 음소의 결합으로 이루어져 있기 때문에 이를 음운 단위의 샘플로 분할하는 텍스트 전처리 과정을 필요로 한다. 한국어 음성합성에서는 자소(grapheme)단위로 분할하는 방법과 분할된 자소 시퀀스를 음소(phoneme) 단위로 변환하는 G2P를 거치는 방법 등이 사용된다.

2.2.1 한국어 자소 분리

한국어 음성합성에서 각 음절이 발음될 때 초성, 중성, 종성의 순서대로 발음하게 된다. 또한, 음운 변화 등의 현상으로 같은 음절도 해당 음절 주변의 다른 음절들의 영향으로 다르게 발음될 수 있다. 자소 시퀀스만을 사용하게 되면 명시적으로 이러한 음운 변화를 반영하지는 못하며 주로 어텐션 가중치의 분산으로 주변 자소의 정보를 반영하여 발음을 생성하게 된다. 이러한 음운 변화를 반영하기 위해 G2P(Grapheme to phoneme)를 사용하여 발음할 텍스트를 발음 규칙에 맞게 변환할 수 있으나, 음운 변화 등의 정보가 미리 반영된 시스템이 필요하게 된다.

2.2.2 자소 연결 정보 분석

음성합성에서 음절내의 특정 자소가 발음될 때에 주변의 자소를 반영하면 보다 자연스러운 발음을 생성할 수 있다. 예를 들어 자음 동화 현상으로 “막는[망는]”에서 종성 ‘ㄱ’은 직후의 초성 ‘ㄴ’의 영향으로 종성 ‘ㅇ’으로 발음하게 된다. G2P를 사용하면 발음 규칙상의 발음 시퀀스로 변환하므로 이를 명시적으로 반영하지만, 자소 시퀀스를 입력으로 사용하게 되면 “막는”을 시퀀스 변환 시 [(초성)ㅁ,(중성)ㅏ,(종성)ㄱ,(초성)ㄴ,(중성)ㅡ,(종성)ㄴ]으로 변환하여 음운 변동 정보를 반영할 수 없게 된다.

이를 해결하기 위해 텍스트 처리 과정에서 각 자소에서 직후 자소 정보를 추가로 반영하여 자소들 사이의 연결 정보가 반영된 시퀀스를 생성할 수 있다. 예를 들어 “막는”을 [(초성)ㅁ,(중성)ㅏ,(종성)ㄱ,(초성)ㄴ,(중성)ㅡ,(종성)ㄴ]으로 변환 후, 각 샘플 사이에 전후 샘플의 정보를 반영하여 [(초성)ㅁ, {ㅁ+ㅏ},(중성)ㅏ, {ㅏ+ㄱ}, (종성)ㄱ, {ㄱ+ㄴ}, (초성)ㄴ, {ㄴ+ㅡ}, (중성)ㅡ, {ㅡ+ㄴ}, (종성)ㄴ]으로 변환한다.

위와 같이 시퀀스 확장을 수행할 경우 연결 정보를 반영한 샘플에서는 인접한 자소에 따라 음운 변동이 일어나게 될 때의 정보를 학습할 수 있게 된다. 자음 동화, 구개음화, 경음화, 자음 축약 등의 음운 변동 현상이 인접한 자음, 모음들 사이에서 발생하는 현상으로 각 현상마다 발음 규칙을 반영한 G2P를 설계하지 않고도 이를 반영하여 자연스러운 발음이 가능해진다. 또한, 음성합성은 텍스트 시퀀스로부터 스펙트로그램 시퀀스를 추정할 때 두 시퀀스 사이의 길이의 불균형이 존재하는데, 이는 시퀀스 투 시퀀스[14] 모델에서 결과 시퀀스 추정 과정에서 부정적인 영향을 끼칠 수 있다. 상대적으로 더 긴 시퀀스인 스펙트로그램을 추정할 때 길이가 상대적으로 짧고 정보량이 적은 텍스트로부터 정보를 생성해야하기 때문이다. 자소 연결 정보를 텍스트 시퀀스에 추가하게 되면 시퀀스 투 시퀀스 모델의 입력 시퀀스가 길어지며 부가적인 정보를 줄 수 있게 되므로 스펙트로그램 추정 시 표현력을 높일 수 있게 된다.

Ⅲ. 실 험

3.1 실험 환경

본 논문에서는 약 10시간 분량의 단일화자 한국어 <텍스트, 음성> 데이터베이스를 사용하여 비교실험을 진행하였다.

멜 스펙트로그램은 80차이며, 주파수 도메인의 신호처리에서 16밀리초의 프레임 단위로 윈도우 크기는 64밀리초의 구성을 사용하였다. 멜 스펙트로그램은 위치 민감성 해싱 어텐션을 사용하기 위해 군집 크기의 2배의 배수로 0-패딩을 수행하였다. 멜 스펙트로그램에서 음성 파형으로 복원하는 보코더는 멜 스펙트로그램으로부터 선형 스펙트로그램의 크기를 선형 신경망을 이용해 추정 후, 그리핀-림 알고리즘을 사용하였다. 학습에 사용한 GPU는 1개의 GeForce RTX 2080Ti로 GPU에 캐시 가능한 최대 메모리를 사용하여 학습을 진행하였다. 학습에 사용한 optimizer는 Adam optimizer[15]로 $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$로 optimizer를 구성하였고, 0.0005의 고정된 learning rate를 사용하였다.

3.2 실험 결과 및 평가

학습 과정에서 필요한 메모리를 측정하고자 한글 텍스트와 멜 스펙트로그램의 길이를 임의로 512, 1024로 고정한 후 메모리를 측정하였다. 또한, 학습 시 단일 GPU에서 최대로 사용 가능한 배치 사이즈로 학습하였고, 사용한 메모리를 측정하였다.

표 1. 학습 시 모델별 메모리 사용량 비교
Table 1. Memory comparison between models during training

Model	Batch size	text length	mel length	memory cached (byte)
Transformer	10	512	1,024	$9,986\times10^6$
Reformer	10	512	1,024	**$5{,}148\times10^6$**

표 1에서 리포머 네트워크를 사용하여 음성합성 모델을 학습할 때 같은 미니 배치 사이즈를 절반에 가까운 메모리를 사용하여 학습이 가능한 것을 확인할 수 있었다.

표 2의 결과에서, 리포머 TTS 모델을 학습할 때 한국어 자소의 연결정보를 사용하여 보다 긴 text 시퀀스를 사용하여 학습을 진행하더라도 보다 큰 batch size를 확보할 수 있음을 확인할 수 있었다.

또한, 한국어 연결 정보를 사용하여 리포머 음성합성 모델을 학습한 것과 연결정보를 사용하지 않은 모델을 사용한 결과를 비교하기 위해서 mean opinion score (MOS)와 comparative MOS (CMOS) test를 수행하였다. 15명의 음성 전문가를 대상으로 선호도 테스트를 진행하였으며, 신뢰구간 95%로 오차구간을 계산하였다. MOS 및 CMOS test 결과는 표 2와 같다.

표 3의 주관적 성능 평가 결과로 보았을 때, 한국어의 연결정보를 사용하여 음성합성을 수행하였을 때 연결정보를 사용하지 않고 한국어의 단일 초성, 중성, 종성 정보만을 사용한 경우보다 성능 향상이 있음을 확인할 수 있었다.

표 2. 모델 별 한국어 연결정보 사용 여부에 따른 batch size 및 메모리 사용량
Table 2. Batch size and memory usage with repect to whether grapheme is linked between models

Model	Batch size	memory usage (MiB)
Transformer w/o text linking	10	10,831
Transformer with text linking	7	9,971
Reformer w/o text linking	16	10,859
Reformer with text linking	14	10,665

표 3. 한국어 연결정보 사용에 따른 음성 선호도 성능 평가
Table 3. Preference evaluation with respect to whether grapheme is linked

Model	MOS	CMOS
Reformer TTS w/o text linking	3.299 ± 0.0585	0
Reformer with text linking	3.769 ± 0.0526	0.628 ± 0.0764

한편, 한국어의 연결정보를 사용하였을 때, 연결정보를 사용하지 않은 경우와 비교하여 어텐션의 경향성을 확인하기 위해 어텐션 궤적을 그래프로 나타내었으며 전체 스택의 어텐션 중 일부의 그래프는 그림 4, 5, 6, 7와 같다. 그림 4, 5, 6은 각각 연결 정보를 사용하지 않고, G2P를 사용하여, 그리고 연결정보를 사용하여 음성합성을 진행하였을 때 텍스트와 스펙트로그램 사이의 정렬이 나타난 그래프로, 연결정보를 사용하거나 G2P를 사용하였을 때, 연결정보를 사용하지 않았을 때보다 어텐션이 텍스트 샘플에 대해 뚜렷하게 나타나는 것을 확인할 수 있었다. G2P와 같이

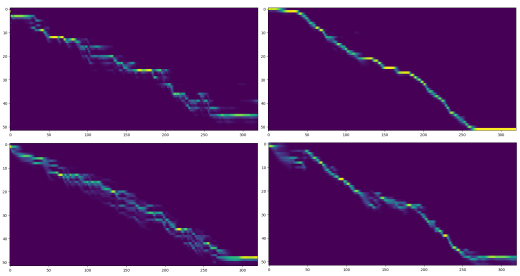

그림 4. 텍스트 연결정보를 사용하지 않은 어텐션 그래프
Fig. 4. Attention graph without text linking

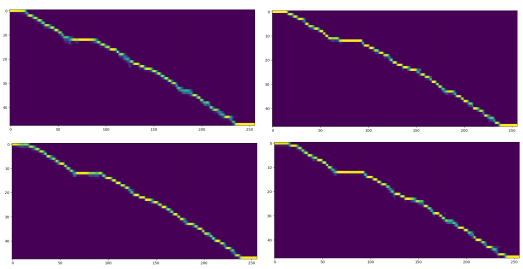

그림 5. G2P를 사용한 어텐션 그래프
Fig. 5. Attention graph with g2p

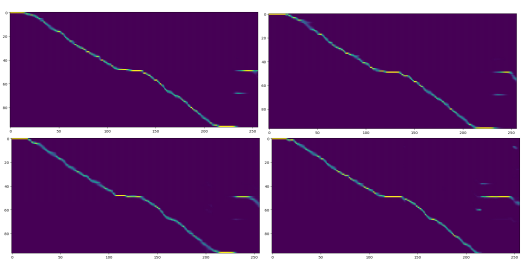

그림 6. 텍스트 연결정보를 사용한 어텐션 그래프
Fig. 6. Attention graph with text linking

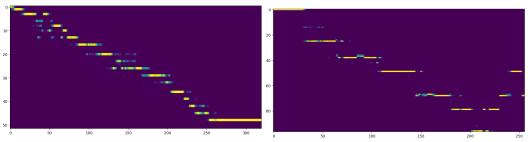

그림 7. 텍스트-스펙트로그램 정렬 정보가 아닌 어텐션 그래프
Fig. 7. Attention without text-spectrogram alignment information

언어학적 지식을 바탕으로 설계된 모델을 사용하지 않더라도, 연결정보를 사용할 경우 어텐션 수행 시 주목할 텍스트를 명확히 결정할 수 있게 된다.

그림 7의 어텐션 궤적은 텍스트와 스펙트로그램의 정렬이 이루어지지 않은 스택의 어텐션 궤적으로, 왼쪽은 연결정보가 사용되지 않은 음성합성 모델, 오른쪽은 연결정보가 사용된 음성합성 모델로부터 얻은 어텐션 그래프이다. 텍스트와 음성 사이의 직접적인 정렬이 이루어지지 않더라도 특정 쿼리 시점에서 곧 등장할 띄어쓰기에 어텐션 값에 가중치가 높은 것을 확인할 수 있었는데, 이는 음성 생성 시 곧 등장할 발음 공백 구간을 모델 내에서 고려하고 있다고 해석할 수 있다.

Ⅳ. 결 론

한국어 음성합성에서 리포머 음성합성 모델을 통해 메모리 측면에서 효율적인 학습을 진행할 수 있고, 효율적인 메모리 사용량만큼 보다 많은 텍스트 정보를 음성합성 입력으로 사용할 수 있게 된다. 또한, 한국어 텍스트의 연결 정보를 사용한 텍스트 정보의 보간을 통해 G2P를 사용하지 않더라도 어텐션 계산 시 주목할 텍스트가 명확히 결정하도록 음성합성 시스템을 구성할 수 있다.

References

[1] A. W. Black, H. Zen, and K. Tokuda, “Statistical parametric speech synthesis,” in *IEEE ICASSP’07, vol. 4, pp. IV-1229-IV-1232,* Honolulu, HI, USA, 2007.

[2] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for hmm-based speech synthesis,” in *Proc. IEEE Int. Conf. Acoust., Speech, and Sign. Process.* (Cat. No.00CH37100), vol. 3, pp. 1315-1318, 2000.

[3] S. O. Arik, et al., “Deep voice: Real-time neural text-to-speech,” in *Proc. 34th Int. Conf. Mach. Learn.,* vol. 70, pp. 195-204, 2017.

[4] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, “Deep voice 3: Scaling text-to-speech with convolutional sequence learning,” in *arXiv preprint arXiv:1710.07654*, 2017.

[5] Y. Wang, et al., “Tacotron: Towards end-to-end speech synthesis,” in *Proc. Interspeech 2017,* pp. 4006-4010, 2017.

[6] J. Shen, et al., “Natural tts synthesis by conditioning wavenet on mel spectrogram predictions,” in *IEEE ICASSP,* pp. 4779-4783, 2018.

[7] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, “Neural speech synthesis with transformer network,” in *Proc. AAAI Conf. Artificial Intell.,* vol. 33, pp. 6706-6713, 2019.

[8] J. Lee, S. Cheon, B. Choi, N. Kim, and D. Hong, “Speech style modeling method using mutual information for end-to-end speech synthesis,“ *J.KKS,* vol. 44, pp. 1641-1647, 2019.

[9] N. Kitaev, L. Kaiser, and A. Levskaya, “Reformer: The efficient transformer,” in *Int. Conf. Learn. Representations*, 2019.

[10] A. Andoni, P. Indyk, T. Laarhoven, I. Razenshteyn, and L. Schmidt, “Practical and optimal lsh for angular distance,” in *Advances in NIPS,* pp. 1225-1233, 2015.

[11] A. N. Gomez, M. Ren, R. Urtasun, and R. B. Grosse, “The reversible residual network: Backpropagation without storing activations,” in *Advances in NIPS,* pp. 2214-2224, 2017.

[12] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE CVPR,* pp. 770-778, 2016.

[13] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, “Forward attention in sequence-to-sequence acoustic modeling for speech synthesis,” in *IEEE ICASSP,* pp. 4789-4793, 2018.

[14] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural

networks," in *Advances in NIPS,* pp. 3104-3112, 2014.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *arXiv preprint arXiv:1412.6980,* 2014.

임 형 래 (Hyeong Rae Ihm)

2018년 2월 : 성균관대학교 전자전기공학부 졸업

2020년 8월 : 서울대학교 전기정보공학부 석사 졸업

<관심분야> 음성 신호 처리, 음성 합성, 뉴럴 네트워크

[ORCID:0000-0002-2687-5724]

천 성 준 (Sung Jun Cheon)

2013년 8월 : 서울대학교 전기정보공학부 학사 졸업

2014년 3월~현재 : 서울대학교 전기정보공학부 석박사통합과정 박사과정

<관심분야> 음성 신호 처리, 음성 합성, 뉴럴 네트워크

[ORCID:0000-0002-7293-6997]

최 병 진 (Byoung Jin Choi)

2013년 5월 : University of Wisconsin-Madison 전기공학부 학사 졸업

2017년 3월~현재 : 서울대학교 전기정보공학부 석박사통합과정 박사과정

<관심분야> 음성 신호 처리, 음성 합성, 뉴럴 네트워크

[ORCID:0000-0003-1319-8215]

김 민 찬 (Min Chan Kim)

2019년 2월 : 고려대학교 전자전기공학부 졸업

2019년 2월~현재 : 서울대학교 전기정보공학부 석박사통합과정 석사과정

<관심분야> 음성 신호 처리, 음성 합성, 뉴럴 네트워크

[ORCID:0000-0002-8150-765X]

김 남 수 (Nam Soo Kim)

1988년 : 서울대학교 전자공학과 학사 졸업

1990년 : 한국과학기술원 전기 및 전자공학과 석사 졸업

1994년 : 한국과학기술원 전기 및 전자공학과 박사 졸업

1993년~1998년 : 삼성종합기술원 전문연구원

1998년~현재 : 서울대학교 전기정보공학부 교수

<관심분야> 음성 신호 처리, 음성 인식, 통계적 신호처리, 패턴 인식, 휴먼 인터페이스

[ORCID:0000-0002-0568-4902]