전이 확률 통제를 활용한 Neural Transducer 기반 음성 합성 시스템의 발화 속도 제어

김민찬, 최병진, 김남수
서울대학교 전기정보공학부 뉴미디어통신공동연구소 휴먼인터페이스 연구실

{mckim, bjchoi}@hi.snu.ac.kr, nkim@snu.ac.kr

Speech Rate Control for Neural Transducer based Text-to-Speech through Transition Probability Restriction

Minchan Kim, Byoung Jin Choi and Nam Soo Kim
Human Interface Laboratory,
Department of Electrical and Computer Engineering and INMC,
Seoul National University

요 약

본 논문은 neural transducer 기반 음성 합성 시스템에서 발화 속도를 조절하는 기법에 관한 연구이다. 본 논문에서는 semantic token 을 활용한 2 stage 음성 합성 시스템에서 front-end 인 neural transducer 모델의 발화 속도를 조절하기 위해 semantic token 생성 시 기학습된 neural transducer 의 전이 확률을 통제하는 방법을 제안한다. 실험을 통해 제안된 알고리즘을 이용해 별도의 학습 없이 효율적인 방법으로 자연스럽게 발화 속도를 조절할 수 있음을 확인하였다.

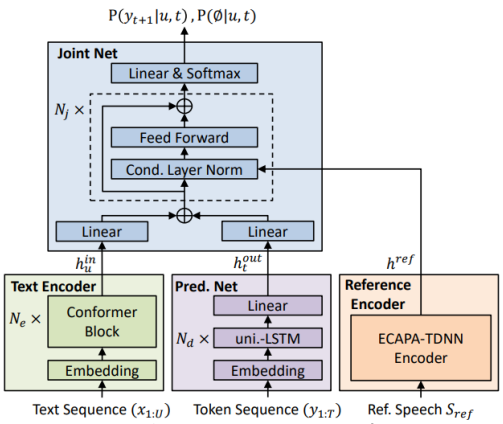
Ⅰ. 서 론

기존의 딥러닝 기반 음성 합성 시스템은 발화 속도(말하는 빠르기)를 조절하는 데에 어려움이 있다. 대표적인 접근 방식으로는 화자 정보나 참조 음성으로부터 추출한 style embedding 을 이용해 원하는 발화 속도를 간접적으로 입력하는 방식이 있고, 그 밖에 duration 기반의 음성 합성 시스템은 각 음소별 duration 을 직접 조정하여 발화 속도를 조절하는 기법 등이 있다. 이 중 style embedding 을 이용한 조절은 추가적인 입력을 필요로 하고 발화 전체의 속도를 조절할 수 있지만 부분별로 세세한 조절이 불가능하다는 단점이 있다. 또한 duration 을 직접 조정하는 방식은 음소별 duration 을 조정하는 추가적인 작업이 필요하고 강제로 조정된 duration 으로 인해 음성이 부자연스러워지는 문제점이 있다.

최근 제안된 neural transducer[1]을 활용한 음성 합성 시스템 [2] 에서는 neural transducer 의 시퀀스 생성 기능을 통해 발화 속도가 결정된다. Neural transducer 은 입력과 출력 사이의 monotonic alignment constraint 를 가정한 sequence-to-sequence(seq2seq) 모델로, 자기 회귀(autoregressive)적인 방식으로 출력을 생성하고 blank token(∅)이 출력될 때 마다 다음 입력 단위로 전이 (transition) 되며 출력을 생성한다. 본 논문에서는 neural transducer 기반의 음성 합성 시스템의 추론 시 전이 확률, 즉 blank token 을 출력할 확률을 조절하여 음소 별 duration 을 자연스럽게 조절하는 기법을 제안한다. 실험을 통해 해당 기법을 활용해 추가적인 통제 입력이나 학습 없이 발화 속도 조절이 가능한 것을 확인하였다.

Ⅱ. 본론

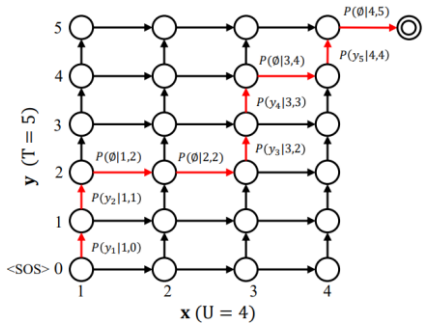
본 연구는 semantic token 을 활용한 2 stage 기반의 음성 합성 시스템 [2]을 기반으로 진행하였다. 이 때 semantic token 은 wav2vec2.0 embedding [3]에 대한 k-means clustering 의 index sequence 로, 해당 시스템은 text 에서 semantic token 으로 변환하는 단계(Reading)와 semantic token 에서 speech 로 변환하는 단계(Speaking)로 구성된다. 이 중 발화 속도와 관련된 문제는 Reading 단계의 token transducer 에 의해 이뤄지게 되고, 이는 neural transducer 구조를 가진다.



[그림 1] Token transducer 의 구조도

해당 token transducer 는 text encoder, prediction network, reference encoder, joint network 로 구성된다. 이 중 joint network 는 text encoder 의 출력과

prediction network 의 출력을 입력 받아 최종적으로 [그림 2]와 같은 transducer lattice 의 emission 확률과 transition 확률을 출력하게 된다. 추가로 reference encoder 는 참조음성을 입력받아 reference embedding 을 출력하고, 이는 joint network 에 컨디셔닝 되어 별도의 속도 조절을 가능하게 한다.

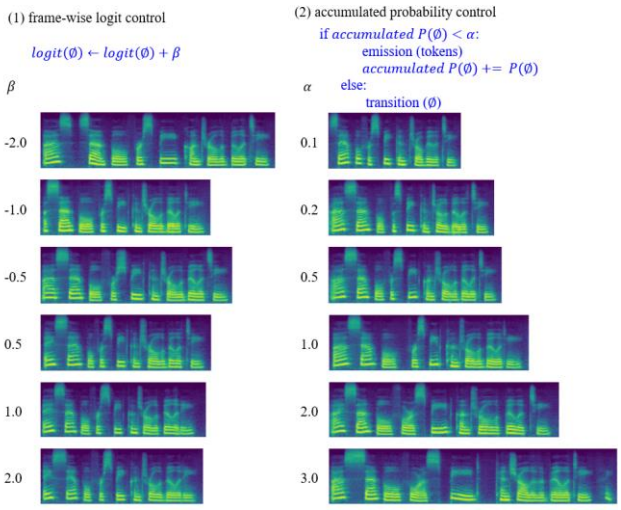


[그림 2] Token transducer 의 alignment lattice

Token transducer 는 생성 시 blank token 을 출력함에 따라 alignment 의 다음 음소 단위로 전이하게 되고, 이를 통해 음소별 duration 을 정의할 수 있다. 이 때 전이 확률에 제약을 둠으로써 각 음소에서의 전이를 늦추거나 앞당겨 duration 을 조절하여 발화를 느리거나 빠르게 통제할 수 있다. 이는 기존의 duration 기반 모델에서 duration 을 강제로 조정하는 것과 같이 생성된 duration 을 조정하는 것이 아니라 자기회귀적으로 동작하는 neural transducer 의 context 에서 duration 을 순차적으로 간접 통제하기 때문에 기존 방식에 비해 자연스러운 음성을 생성할 수 있다.

전이 확률 통제에는 여러 방식을 적용해볼 수 있다. (1) 우선 생성 시 joint network 에 의해 계산된 전이 확률의 logit 값에 bias 를 더해주는 방식이 있다. 이는 전이 확률에 상수를 곱한 뒤 정규화해주는 것과 유사한 의미를 갖는다. 이러한 방식은 현재의 전이 확률에 대해서만 통제 하는 방식으로 볼 수 있다. (2) 또는 음소별로 전이 확률의 누적 값에 threshold 를 두고 이를 넘길 경우 전이시키는 방식이 있다. 이는 전이를 샘플링 기반이 아닌 누적 확률 값을 기준으로 하기 때문에 조금 더 안정적인 생성이 가능하다.

실험을 위해 token transducer 에 (1)과 (2)의 기법을 적용해보았다. 추가적으로 발화 속도를 제어할 수 있는 참조 음성과 텍스트를 고정한 후 각 기법의 parameter 를 조정해가며 음성을 생성해보았고, 결과는 아래와 같다.



[그림 3] parameter 에 따른 생성된 음성의 발화 속도

[그림 3]의 결과에 따르면 전이 확률의 logit 값에 bias 를 더하는 기법은 bias 가 증가함에 따라 transition 이 더욱 빈번하게 발생하여 발화 속도가 빨라지는 것을 기대할 수 있으나 실험 결과 규칙적인 양상이 나타나지 않는 것을 확인할 수 있었다. 이는 샘플링 기반의 자기 회귀 방식 모델의 출력 확률에 가중치를 두는 것이 현재의 확률 값에 지배적으로 영향을 받기 때문인 것으로 추정할 수 있다. 달리 말해 현재의 전이 확률 값이 매우 낮을 경우 가중치의 영향을 거의 받지 않고 가중치의 영향을 받기 위해서는 이미 높은 전이 확률을 가지게 된다. 반면 누적된 전이 확률에 threshold 를 두는 경우 threshold 를 크게 둠에 따라 안정적으로 발화 속도가 느려지는 양상을 확인할 수 있다. 뿐만 아니라 발화 시 parameter 값에 따른 발음 손실 등의 문제점도 앞선 방식에 비해 우수한 것을 확인할 수 있었다. 해당 실험을 통해 제안된 neural transducer 기반 모델의 발화 속도 조절 알고리즘을 활용해 세밀한 제어를 하기는 힘들지만 전반적인 모델의 발화 속도를 제어할 수 있는 것으로 기대해볼 수 있다.

Ⅲ. 결론

본 논문에서는 neural transducer 기반의 음성 합성 시스템에서 전이 확률에 제약을 두는 방식을 통해 발화 속도를 제어하는 기법을 제안한다. 해당 기법을 통해 기존의 기법과 달리 음소별 duration 을 직접 조정하지 않고 context 에 맞게 간접적으로 제어하는 방식을 통해 자연스러운 통제가 가능하다. 실험을 통해 전이 확률의 누적 값에 제약을 둘 때 발화 속도 조절이 안정적으로 동작하는 것을 확인하였다.

ACKNOWLEDGMENT

이 논문은 2023 년도 BK21 FOUR 정보기술 미래인재 교육연구단에 의하여 지원되었음.

참 고 문 헌

[1] Graves, Alex. "Sequence transduction with recurrent neural networks." *arXiv preprint arXiv:1211.3711* (2012).

[2] Kim, Minchan, et al. "Transduce and Speak: Neural Transducer for Text-to-Speech with Semantic Token Prediction." *arXiv preprint arXiv:2311.02898* (2023).

[3] Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations." *Advances in neural information processing systems* 33 (2020): 12449-12460.