비스트리밍 음성 합성 모델의 세그먼트 기반 실시간 추론 기법

박재홍, 전용현, 김남수
서울대학교 전기정보공학부 뉴미디어통신공동연구소

{jhpark, yhjeon}@hi.snu.ac.kr, nkim@snu.ac.kr

Real-Time Speech Synthesis via Segment-Based Inference on Non-Streaming Models

Jaehong Park, Yong Hyeon Jun, Nam Soo Kim
Department of Electrical and Computer Engineering and INMC, Seoul National Univ.

요 약

본 논문에서는 스트리밍 방식으로 학습되지 않은 음성 합성 모델을 실시간 환경에 적용하기 위한 세그먼트 기반 추론 기법을 제안한다. 입력 텍스트를 구간별로 분할한 뒤 각 세그먼트에 대해 순차적으로 음성을 합성하고, 생성된 음성 구간을 후처리를 통해 연결함으로써 연속적인 출력을 구성한다. 구체적으로, 단순 연결(concatenation) 방식과 함께 linear 및 Hann window 를 활용한 overlap-add(OLA) 기법을 적용하여 세그먼트 경계에서 발생하는 불연속성을 완화한다. 또한 모델의 receptive field 를 고려한 버퍼를 도입하여 프레임을 연속적으로 생성함으로써 세그먼트 경계에서의 불연속성을 최소화한다. 제안한 방법은 추가적인 스트리밍 학습 없이도 기존 비스트리밍 음성 합성 모델을 실시간 합성 시나리오에 효과적으로 적용할 수 있음을 보여준다.

Ⅰ. 서 론

실시간 음성 합성은 대화형 시스템 및 스트리밍 음성 응용에서 중요한 요구사항으로 자리 잡고 있다. 이를 위해 스트리밍을 고려한 모델 구조나 학습 기법이 제안되고 있으나, 이러한 접근은 모델 설계 및 학습 과정에 추가적인 복잡성을 수반하며, 기존에 비스트리밍 환경에서 학습된 음성 합성 모델을 그대로 활용하기 어렵다는 한계를 가진다.

한편, 많은 음성 합성 모델은 전체 문장 또는 충분한 길이의 입력을 가정하고 학습되며, 추론 과정 또한 비스트리밍 환경을 전제로 설계되어 있다. 이러한 모델을 실시간 환경에 적용할 경우 입력 지연, 세그먼트 경계에서의 불연속성, 음질 저하 등의 문제가 발생할 수 있다. 그럼에도 불구하고, 기존 비스트리밍 모델이 제한적인 조건 하에서 실시간 합성에 활용될 수 있는지에 대한 체계적인 검토는 상대적으로 부족한 상황이다.

본 논문에서는 입력 텍스트를 세그먼트 단위로 분할하여 순차적으로 음성을 생성하는 단순한 추론 전략을 기반으로, 비스트리밍 음성 합성 모델의 실시간 적용 가능성을 실험적으로 분석한다. 생성된 음성 세그먼트는 단순 연결(concatenation) 방식 또는 linear 및 Hann window 를 활용한 overlap-add(OLA) 기법을 통해 결합되며, 또한 모델의 receptive field 를 고려한 버퍼링을 통해 프레임 단위의 연속 생성 가능성을 확인한다.

본 연구의 목적은 추가적인 스트리밍 학습이나 모델 구조 변경 없이도 비스트리밍 음성 합성 모델이 실시간 합성 시나리오에 어느 정도까지 적용 가능할 수 있는지를 검증하는 데 있다. 이를 통해 실시간 음성 합성을 위한 실용적인 설계 선택지와 그 한계를 명확히 제시하고자 한다.

Ⅱ. 본론

1) SegINR 기반 음성 표현

본 연구에서 사용한 음성 합성 모델은 text encoder, segment-wise implicit neural representation(SegINR), 그리고 vocoder 로 구성된다. 이 중 SegINR 은 텍스트 입력과 음성 간의 정렬을 담당하는 핵심 모듈로, text embedding 으로부터 각 토큰에 대응하는 duration 을 내부적으로 추정하고 이를 기반으로 세그먼트 단위의 음성 표현을 생성한다. 이를 통해 프레임 단위 시퀀스 모델에 의존하지 않고도 음성 합성이 가능하다.

텍스트 인코더를 통해 추출된 토큰 단위 표현은 SegINR 에 입력되며, SegINR 은 각 토큰에 대응하는 세그먼트별 시멘틱 토큰을 출력한다. 생성된 시멘틱 토큰은 세그먼트 단위의 음성 정보를 압축적으로 표현하며, 이후 vocoder 의 입력으로 사용되어 최종 음성이 합성된다. 이러한 구조를 통해 텍스트-음성 간의 길이 정렬은 별도의 duration predictor 없이 모델 내부에서 처리되며, 세그먼트 단위의 표현은 입력 텍스트를 구간별로 처리하는 순차적 추론 방식과도 자연스럽게 결합된다.

또한 해당 구조는 비스트리밍 환경에서 학습된 모델이라 하더라도 입력 텍스트를 순차적으로 처리하며 음성을 생성할 수 있도록 하여, 실시간 합성 시나리오에의 적용 가능성을 제공한다.

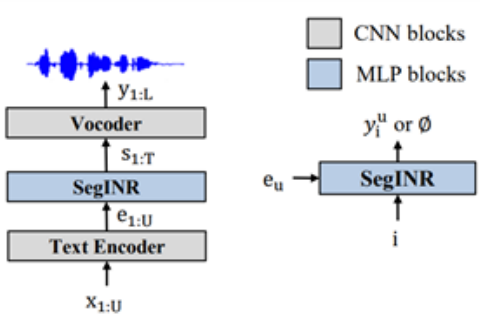

그림 1. SegINR 기반 음성 합성 모델의 시스템 도식

2) Overlap-Add 기반 세그먼트 연결

세그먼트 단위로 생성된 음성 또는 음향 특징을 연결하기 위해, 본 연구에서는 overlap-add(OLA) 기반의 후처리 방식을 고려한다. OLA 는 인접한 세그먼트 간에 일정 구간을 중첩시킨 뒤 가중 합을 수행함으로써, 경계 부근에서 발생할 수 있는 불연속성을 완화하는 데 널리 사용되는 방법이다. 세그먼트 단위 합성 환경에서는 각 구간이 독립적으로 생성되기 때문에, 이러한 중첩 기반 연결 방식이 음성의 연속성을 유지하는 데 효과적일 수 있다.

본 연구에서는 OLA 적용 시 단순한 선형 가중치(linear window)와 Hann window 를 사용한 가중 방식을 검토한다. Linear window 는 중첩 구간에서 가중치를 선형적으로 변화시켜 두 세그먼트를 부드럽게 전환하며, Hann window 는 경계에서의 에너지 변화를 완만하게 조절하여 보다 매끄러운 연결을 유도한다. 이러한 가중 방식들은 세그먼트 경계에서의 음질 열화나 클릭 노이즈를 줄이기 위한 일반적인 선택지로, 본 연구에서는 세그먼트 기반 실시간 추론 과정에서의 적용 가능성을 중심으로 활용한다.

3) 실험 및 결과

실험은 실시간 합성 시나리오에 적용하는 두 가지 추론 방식을 비스트리밍 음성 합성 모델에 적용하여 진행하였다. 첫 번째는 세그먼트 단위 합성 후 연결 방식이며, 두 번째는 버퍼를 이용한 프레임 단위 연속적 생성 방식이다. 모든 실험은 동일한 모델 구조를 사용하여 수행되었으며, 추론 방식에 따른 차이를 비교하였다.

세그먼트 단위 합성 실험에서는 입력 텍스트의 토큰들을 일정 길이의 구간으로 분할한 뒤, 각 구간에 대해 독립적으로 음성 합성을 수행하고 생성된 결과를 연결하였다. 세그먼트 간 연결 방식은 세 가지로, concatenation, linear OLA, 그리고 Hann OLA 이다. Concatenation 에서는 15 토큰을 하나의 세그먼트로 하여 세그먼트 단위 음성을 생성한 뒤, 별도의 후처리 없이 단순 연결을 하였다. Linear OLA 와 Hann OLA 에서는 20 토큰을 하나의 세그먼트로 하여 5 개 토큰에 대한 부분을 중첩시켜 overlap-add 방식을 적용하였다.

버퍼를 이용한 프레임 단위 합성 실험에서는 모델 내부 레이어들의 receptive field 를 고려한 버퍼를 도입하여 음성을 프레임 단위로 순차적으로 생성하였다. 이 방식은 세그먼트 경계에서의 불연속성을 완화하며, 기존의 비스트리밍 시나리오와 유사한 결과물을 만들어내는 것을 목표로 한다.

표 1. 합성된 음성의 음질 평가

방식 구분		MOS ↑
Baseline		3.4
세그먼트 단위 합성	Concatenation	2.2
	Linear OLA	3.0
	Hann OLA	2.4
프레임 단위 합성		3.4

세그먼트 단위 합성 방식에서 단순 concatenation 은 가장 낮은 성능을 보였으며, 이는 세그먼트 경계에서 발생하는 불연속성과 클릭 노이즈의 영향으로 해석된다. Overlap-add 기법을 적용한 경우 음질이 전반적으로 개선되었으며, 특히 linear OLA 는 concatenation 대비 유의미한 향상을 보였다. 버퍼를 이용한 프레임 단위 연속적 생성 방식은 baseline 과 동일한 성능을 달성하였다. 이는 비스트리밍 환경에서 학습된 모델이라 하더라도, 프레임 레벨의 연속적 생성이 음질 측면에서 효과적인 대안이 될 수 있음을 보여준다.

표 2. Inference speed 측정 결과

방식 구분		RTF ↓
Baseline		0.0323
세그먼트 단위 합성	Concatenation	0.0487
	Linear OLA	0.0624
	Hann OLA	0.0664
프레임 단위 합성		0.4267

RTF 측면에서 baseline 은 가장 빠른 추론 속도를 보였다. 세그먼트 단위 합성 방식은 연결 및 후처리로 인한 추가 연산으로 baseline 대비 RTF 가 증가하였으며, concatenation 이 가장 낮은 RTF 를 기록하였다. OLA 를 적용한 경우 중첩 및 가중 연산으로 인해 RTF 가 더 증가하였다. 버퍼를 이용한 프레임 단위 합성 방식은 가장 높은 값을 기록하여, 음질 측면의 장점과 함께 계산 비용 증가라는 trade-off 가 존재함을 확인하였다.

Ⅲ. 결론

본 연구에서는 비스트리밍 환경에서 학습된 음성 합성 모델을 대상으로, 세그먼트 단위 합성 및 버퍼 기반 프레임 단위 생성 방식을 통해 실시간 적용 가능성을 분석하였다. 실험 결과, 세그먼트 기반 방식은 경계 처리 기법에 따라 음질과 속도 간의 trade-off 를 보였으며, 버퍼를 이용한 프레임 단위 생성은 baseline 에 근접한 음질을 유지할 수 있음을 확인하였다. 이는 추가적인 스트리밍 학습 없이도 기존 음성 합성 모델을 실시간 합성 시나리오에 적용할 수 있는 실용적인 설계 방향을 제시한다.

ACKNOWLEDGMENT

이 논문은 2026 년도 BK21 FOUR 정보기술 미래인재 교육연구단에 의해 지원되었음.