한글의 음절 단위 구분성을 이용한 한국어 음성 합성

박재홍, 문성환, 김남수
서울대학교 전기정보공학부 뉴미디어통신공동연구소

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

Korean speech synthesis using syllable-level distinction of Hangul

Jaehong Park, Sung Hwan Mun, Nam Soo Kim
Department of Electrical and Computer Engineering and INMC, Seoul National Univ.

요 약

각 언어와 문자는 저마다의 특성을 갖고 있으므로 음성 합성 모델에서 데이터 전처리를 하는 방식도 그에 따라 달라져야 한다. 한글은 자음과 모음으로 이루어진 음소 문자이며 이러한 음소가 두 개 혹은 세 개가 하나의 글자를 이루기 때문에 음절의 구분이 명확하다. 본 연구에서는 이러한 한글의 특성을 반영한 text encoder 를 이용해 좋은 품질의 한국어 음성 합성을 위한 텍스트 데이터 전처리 방식을 찾고자 하였다. 그 방법으로 텍스트를 embedding 으로 변환하기 전에 음절 단위의 구분 기호를 추가하는 방식을 제안하였고 그 결과로 자연스럽고 정확한 발음을 가진 음성을 합성하였다.

Ⅰ. 서 론

TTS(Text-to-Speech)는 입력된 텍스트에 해당하는 음성을 합성하는 모델로 음성 안내, 번역기, 인터넷 방송 등 다양한 분야에 활용되고 있다. 음성 합성 모델의 학습은 음성과 텍스트 데이터의 전처리, encoder 와 decoder 를 거쳐 합성된 음성의 테스트로 이루어지고 최종적으로 자연스러우며 주어진 텍스트를 명확하게 발성하는 음성의 합성을 목적으로 한다. 본 연구에서는 좋은 품질의 음성을 합성하기 위한 데이터 전처리 방식을 찾는 것을 목표로 했으며, 특히 텍스트 데이터 전처리에 초점을 두었다. 텍스트 데이터 전처리 방법에는 소문자화, 줄임말 확장, phonemize, 빈칸 처리 등 여러가지가 있지만 언어마다 글자의 형태와 나열 방식에 차이가 존재하므로 전처리 과정에는 학습에 사용되는 데이터의 언어에 따라서 차이가 존재하며 본 연구는 한국어와 한글 데이터를 대상으로 하였다.

한글은 자음과 모음의 음소들로 이루어진 음소 문자로 음소의 수가 많지 않아 학습에 용이하며 어절과 음절의 구분이 명확해 텍스트 처리에도 이점을 가진다. 본 연구에서는 이러한 한글의 특성을 이용하여 기존에 영어를 대상으로 연구가 이루어졌던 음성 합성 모델인 VITS(Variational Inference with adversarial learning for end-to-end Text-to-Speech) [1]를 바탕으로 한글 텍스트의 효과적인 전처리 방식을 제안하였다. 그 결과로 합성된 음성의 자연스러움 정도와 발음의 정확성을 MOS 측정을 통해 비교하여 제안된 방식의 효과를 검증하였고 더 좋은 품질의 한국어 음성 합성을 위한 추후 연구 방향을 제시하였다.

Ⅱ. 본론

1) VITS

본 연구에서는 VITS 모델을 기반으로 실험을 진행하였다. VITS 모델은 encoder, monotonic alignment, stochastic duration predictor, decoder 로 구성되며 encoder 부터 decoder 까지 모든 모듈들이 한번에 학습되는 end-to-end 음성 합성 모델이다. 본 연구에서는 encoder 에 들어가기 전 단계인 데이터 전처리 단계에서의 새로운 방식을 제안하여 기존 방식과의 비교를 진행하였다.

VITS 논문의 저자가 연구에 사용한 코드를 공개했으며 해당 코드에서 텍스트 데이터 전처리 부분을 일부 수정하여 사용하였다 [2].

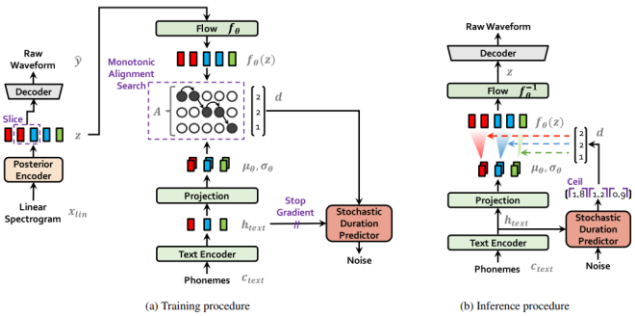

그림 1. VITS 모델의 시스템 도식

2) Korean text encoder

본 연구에서는 한글 데이터를 잘 학습할 수 있도록 설계된 Korean text encoder 를 만들었으며 데이터 전처리를 하는 방식을 두 가지로 하여 품질이 좋은 음성을 합성하는 방식을 탐구하였다. 기존의 encoder 는 영어 텍스트를 입력으로 받아 먼저 아스키 코드로 변환을 하는데, 한글은 아스키 코드로 표현이 불가능하기 때문에 한글 텍스트를 embedding sequence 로 변환하는 새로운 text encoder 가 필요하며 이를 설계함에 있어 고려한 점은 두 가지이다.

첫째는 한글 텍스트를 embedding sequence 로 변환하는 방식으로, 한글을 음절 또는 음소 단위로 변환할 수 있다. 전자는 텍스트에 추가적인 처리가 불필요하고 변환된 embedding sequence 가 짧다는 장점이 있지만 한글에서 나타나는 음절이 총 11,172 개로 매우 많기 때문에 embedding dictionary 가 너무 커진다는 단점이 있다. 이러한 단점과 본 연구의 목표를 고려하였을 때 음소 단위로 더 세밀하게 학습을 시키는 것이 나을 것으로 판단하여 후자의 방식으로 Korean text encoder 를 설계하였다.

둘째는 음절 단위의 구분 방식이다. 한글은 자음, 모음, 받침으로 구성되어 음절의 구분이 명확하기 때문에 한글 텍스트를 음소 단위로 처리함과 동시에 음절 단위의 구분 기호를 삽입하기 매우 쉽다. 이에 착안하여 본 연구에서는 한글 텍스트를 변환함에 있어 추가적인 구분 기호 없이 음소 단위로 변환하는 방식과 음절 단위의 구분 기호를 추가하여 음소 단위로 변환하는 방식 두 가지를 각각 방식 1, 방식 2 로 설정하고 실험을 진행하였다. 각 방식은 다음과 같다.

텍스트 데이터: 안녕하세요?
방식 1: ㅇㅏㄴㄴㅕㅇㅎㅏㅅㅔㅇㅛ?
방식 2: ㅇㅏㄴ/ㄴㅕ/ㅇ/ㅎㅏ/ㅅㅔ/ㅇㅛ/?/

3) 실험 및 결과

실험은 총 두 가지를 진행했으며, Ⅲ에서 제시한 두 가지 방식으로 VITS 모델을 학습시킨 후 각각의 결과물들의 MOS 측정을 진행했다. 이 때 학습에 사용한 데이터셋은 여성 단일 화자의 한국어 음성과 그에 따른 텍스트로 구성된 KSS 데이터셋이다. 첫번째 실험에서는 합성된 음성들의 자연스러움의 정도를 측정했고 방식 1 과 방식 2 로 800,000, 1,600,000, 2,400,000, 3,200,000, 4,000,000 iterations 학습된 모델로 각각 5 개의 문장들에 대한 음성을 합성해 총 50 개의 음성 데이터에 대해 측정을 했다.

표 1. 자연스러움 정도의 MOS 측정 결과

방식 구분	학습 횟수 (x100,000 iterations)	MOS	CI
방식 1	8	3.23	0.17
	16	3.70	0.18
	24	3.83	0.15
	32	3.76	0.20
	40	3.85	0.18
방식 2	8	**3.57**	0.18
	16	**3.96**	0.19
	24	**4.07**	0.17
	32	**4.20**	0.26
	40	**4.15**	0.20

첫번째 실험의 결과로 방식 2 의 MOS 값이 더 높았고 그 이유가 음절 단위의 구분 기호로 인한 발음의 정확성 향상일 것으로 예상되어 두번째 실험으로는 발음의 정확성에 초점을 두어 MOS 측정을 진행했다. 두번째 실험에서는 방식 1 과 방식 2 로 4,000,000 iterations 학습된 모델로 각각 5 개의 문장들에 대한 음성을 합성해 총 10 개의 음성 데이터에 대해 측정을 했다.

표 2. 발음의 정확성의 MOS 측정 결과

방식 구분	MOS	CI
방식 1	3.97	0.16
방식 2	**4.54**	0.19

두번째 실험의 결과로 방식 2 의 MOS 값이 더 높았고 첫번째 실험의 4,000,000 iterations 에서의 MOS 값의 차이인 0.30 보다 두번째 실험의 동일 iterations 에서의 MOS 값의 차이가 0.57 로 더 크게 나타나며 예상대로 발음의 정확성 향상이 자연스러움 정도의 향상에 긍정적인 기여를 했음을 확인하였다.

Ⅲ. 결론

본 연구에서는 한글의 음절 단위 구분이 명확하다는 특징을 이용하여 한국어 음성 합성을 위한 한글 전처리 방식을 제안하였다. 한글을 단순히 음소 단위로 나누는 기존의 전처리 방식과 비교하여 본 방식은 음절 단위로 구분 기호를 추가하였다. 제안한 방식의 성능 측정을 위해 합성된 결과물의 자연스러움 정도와 발음의 정확성의 MOS 측정을 진행했고 그 결과로 제안된 방식이 합성된 음성의 자연스러움 및 발음의 정확성을 향상시키는 것을 확인하였다.

한편 본 연구에서 측정된 자연스러움 정도의 MOS 가 VITS 논문에서 제시한 MOS 보다 낮았는데, 이는 VITS 논문에서 사용한 LJS 데이터셋의 크기와 본 연구에서 사용한 KSS 데이터셋의 크기의 차이, 그리고 학습 시간의 차이로 인한 것으로 생각된다.

마지막으로 합성된 음성의 품질을 높이기 위한 방법으로는 음운론을 기반으로 한 텍스트 전처리로 발음의 정확성을 개선하거나 문장 부호의 추가적인 처리를 통해 문장의 끝 처리를 자연스럽게 하는 등의 방법이 있을 것이다. 그리고 최근에 활발히 진행되고 있는 감정이 담긴 음성 합성 연구를 통해 더욱 자연스러운 음성 합성이 가능할 것이다.

ACKNOWLEDGMENT

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

참 고 문 헌

[1] J. Kim, J. Kong, and J. Son, “Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech,” in *Proc. International Conference on Machine Learning*. PMLR, 2021.

[2] VITS: Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech[Website]. (2023, Sep 25). https://github.com/jaywalnut310/vits