2024년도 한국통신학회 하계종합학술발표회

한국어 음성 합성을 위한 Semantic Token 최적화에 관한 연구

김세민, 김민찬, 정명훈, 김남수

서울대학교 전기정보 공학부 뉴미디어통신공동연구소 휴먼인터페이스 연구실

{smkim21, mckim, mhjeong}@hi.snu.ac.kr, nkim@snu.ac.kr

A Study on Optimization of Semantic Token for Korean Speech Synthesis

Semin Kim, Minchan Kim, Myeonghun Jeong, and Nam Soo Kim

Human Interface Laboratory,

Department of Electrical and Computer Engineering and INMC,

Seoul National University

요 약

본 논문은 semantic tokenization 을 최적화하여 한국어에 대해 optimized 된 발음 정보를 모델링하는 semantic token 을 찾는 연구를 진행하였다. 최근의 language 모델 기반의 음성 합성 연구는 discrete 한 semantic token 과 acoustic token 을 feature 로 활용하여 text-to-semantic token, semantic-to-acoustic token 을 각각 모델링하는 방식으로 진행된다. 따라서, 이때 어떤 semantic token 을 기반으로 학습을 하는가가 전체 프로세스에 큰 영향을 준다. 본 논문에서는 실험을 통해 한국어에 fine-tuned 된 wav2vec embedding 을 사용하는 것이 효과적임을 보였다.

Ⅰ. 서 론

본 논문은 한국어 음성 합성에 효과적인 semantic token 을 찾는 연구를 진행하고 이를 semantic token 을 제시하였다. 최근의 language 모델 기반의 음성 합성 연구 [1, 2]는 discrete 한 semantic token 과 acoustic token 을 feature 로 사용하는 two-stage 형태로 이루어진다. 이 때 효과적인 semantic token 을 사용하는 것이 발음 정보를 모델링함에 있어 유의미한 차이를 보인다. 본 논문에서는 한국어에 fine-tuned 된 wav2vec embedding 을 사용하는 것이 다른 semantic token 에 비해 효과적임을 실험을 통해 보였다.

Ⅱ. 본 론

최근의 Language model 기반의 음성 합성 모델은 discrete 한 semantic token 과 acoustic token 을 intermediate representation 을 활용하여 text-to-semantic token, semantic-to-acoustic token 의 two-stage 구조로 이루어진다. 이 때 semantic token 의 경우 발음 정보를 포함한 phonetic information 을 포함하고 있다고 가정하며, acoustic token 의 경우 좀더 speech 에 가까운 information 을 포함한다고 가정한다.

기존의 semantic token 을 이용한 연구들은 wav2vec2.0[3], Hubert[4]와 같은 self-supervised learning 을 통해 학습한 모델들을 활용하여 embedding 을 추출하고, 이를 k-means cluster 와 같은 clustering 기법을 통해 512, 1024 에 해당하는 discrete token 으로 모델링하였다.

Acoustic token 의 경우 Encodec[5], SoundStream[6] 등의 neural audio codec 을 통해 학습되어 바로 음성으로 변환할 수 있는 codec embedding 을 discrete 한 token 으로 사용하여 음성적 정보를 포함하도록 하였다.

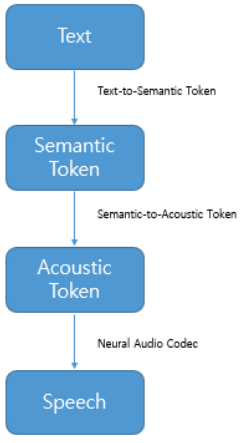

그림 1. Language 모델 기반 음성 합성 모델 구조

본 논문에서는 이 중 기존과 다른 언어를 사용했을 때 가장 큰 차이가 생기는 발음 정보를 모델링하는 semantic token 에 집중하여 한국어 음성 합성에서 더 효과적인 semantic token 을 연구하였다.

Semantic token 이 발음 정보를 제대로 모델링하는지 확인하기 위해 기존의 language model 구조를 사용하기 어려운데, 이는 two-stage 로 이루어진 모델 구조에서는 발음에서 나타나는 오류를 semantic token 만의 문제로 보기 어렵기 때문이다. 따라서, 본 논문에서는 VITS[7]의 구조를 활용하여 semantic token-to-speech 모델을 구성하여 학습하는 방식으로 semantic token 의 효과를 확인하였다.

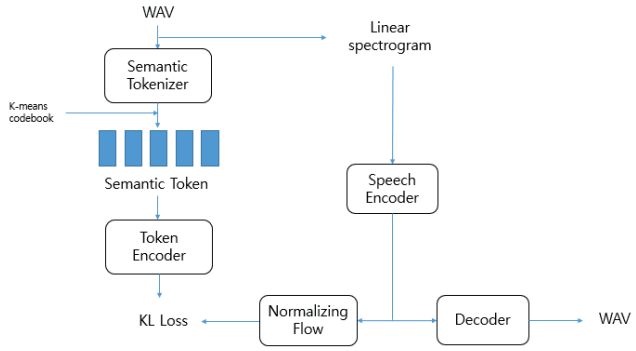

그림 2. Semantic token-to-speech 모델 구조

실제 모델의 학습과정은 다음과 같다. 먼저 wave 를 semantic tokenizer 를 통해 embedding 을 추출하고, 이를 k-means cluster 를 통해 semantic token 으로 변환한다. 이 semantic token 은 token encoder 를 통해 latent representation 으로 변환된다. 또한 같은 wave 를 linear spectrogram 으로 변환하고, 이를 speech encoder 에 통과시켜 latent representation 으로 변환한다. 이를 normalizing flow 를 통해 token encoder 을 representation 과의 KL-divergence 를 줄이는 방향으로 학습한다. 또한 speech encoder 를 통과한 latent representation 이 decoder 를 통과하여 만들어진 wave 를 원래의 wave 와의 L2 loss 를 통해 reconstruction 이 잘 되도록 유도한다.

실험은 내부 남녀 각 1 인인 한국어 음성 합성 데이터셋을 사용하여 진행되었다. Training 에는 약 24 시간의 여성 음성, 약 25 시간의 남성 음성을 사용하였다. Test 로는 약 20 분의 남녀 음성 데이터를 사용하였다. 평가는 발음 정보가 잘 모델링 되는지 확인하기 위해 Test 데이터에 해당하는 오디오 샘플들을 생성한 뒤에 WER 을 측정하여 비교하였다.

베이스라인으로는 기존 모델들이 주로 semantic token 으로 변환할 때 사용하는 wav2vec2.0 을 사용하였다. 또한 더 나은 발음 정보를 모델링하기 위해 wav2vec2.0 을 한국어 데이터에 fine-tuning 한 후에 이를 이용하여 새로운 semantic token 을 추출하고 이를 이용하여 학습하였다.

Model	WER
Wav2vec2.0	4.94
Proposed	4.80

표 1. 베이스 라인 모델과 제시한 모델의 WER

Ⅲ. 결론

본 논문에서는 한국어에서 최적화된 semantic token 을 찾기 위해 semantic token-to-speech 모델을 구성하여 semantic token 의 발음 정보의 모델링 정도를 확인하고, 한국어에 semi-supervised learning 모델을 fine-tuning 하여 더 나은 발음 정보를 모델링할 수 있음을 실험을 통해 보였다.

ACKNOWLEDGMENT

이 논문은 2024 년도 BK21 FOUR 정보기술 미래인재 교육연구단에 의하여 지원되었음.

참 고 문 헌

[1] Kharitonov, Eugene, et al. "Speak, read and prompt: High-fidelity text-to-speech with minimal supervision." Transactions of the Association for Computational Linguistics 11 (2023): 1703-1718.

[2] Borsos, Zalán, et al. "Soundstorm: Efficient parallel audio generation." arXiv preprint arXiv:2305.09636 (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.

[4] Hsu, Wei-Ning, et al. "Hubert: Self-supervised speech representation learning by masked prediction of hidden units." IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021): 3451-3460.

[5] Défossez, Alexandre, et al. "High fidelity neural audio compression." arXiv preprint arXiv:2210.13438 (2022).

[6] Zeghidour, Neil, et al. "Soundstream: An end-to-end neural audio codec." IEEE/ACM Transactions on Audio, Speech, and Language Processing 30 (2021): 495-507.

[7] Kim, Jaehyeon, Jungil Kong, and Juhee Son. "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech." International Conference on Machine Learning. PMLR, 2021.