2023년도 한국통신학회 하계종합학술발표회

Diffusion 생성 모델을 통한 한국어 가창 음성 합성에 관한 연구

김세민, 이현승, 정명훈, 박재홍, 김남수

서울대학교 전기정보 공학부 뉴미디어통신공동연구소 휴먼인터페이스 연구실

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

A Study on the Korean Singing Voice Synthesis via Diffusion generative model

Semin Kim, Hyeon Seung Lee, Myeonghun Jeong, Jaehong Park, and Nam Soo Kim

Human Interface Laboratory,

Department of Electrical and Computer Engineering and INMC,

Seoul National University

요 약

본 논문은 딥러닝 기반의 생성 모델의 하나인 Diffusion 생성 모델을 이용하여 한국어 가창 음성 합성에 대한 연구를 진행하였다. Diffusion 생성 모델은 최근에 발표된 생성 모델로 기존의 생성 모델들에 비해 높은 수준의 샘플을 생성할 수 있는 장점을 가지고 있다. 본 논문에서는 Diffusion 생성 모델을 가창 음성 합성에 적용하고 이를 통해 생성한 가창 음성이 기존의 모델에 비해 질적인 면에서 우수하다는 것을 실험을 통해 확인하였다.

Ⅰ. 서 론

본 논문은 Diffusion 생성 모델을 기반으로 한 한국어 가창 음성 합성 기술을 제시하였다. 가창 음성 합성은 주어진 음과 가사에 따라 적절하고 자연스러운 가창 음성을 합성하는 것을 목표로 한다. 기존의 가창 음성 합성 모델은 MLP-mixer [1], GAN [2] 등을 기반으로 하고 있다. 그러나 이들은 합성 음성의 품질의 한계가 명확하고, 부자연스러운 가창 음성을 생성하는 것에 그쳤다. 최근에 제시된 Diffusion 모델의 경우 보다 높은 품질의 생성을 가능하게 한다는 것이 알려져 있어 이를 가창 음성 합성에 적용하고, 실험을 통해 그 우수성을 확인하였다.

Ⅱ. 본 론

Diffusion 생성 모델 [3], [4]은 다른 생성 모델들과 같이 학습 데이터와 유사한 데이터를 생성하는 것을 목표로 한다. 학습과정에서 Diffusion 모델은 학습 데이터에 노이즈를 점차적으로 더하고 이를 다시 되돌리는 과정을 neural network 를 통해 학습한다. 충분히 학습된 모델은 무작위 노이즈에서 원하는 종류의 데이터를 생성할 수 있게 된다. Diffusion 생성 모델은 forward 와 backward 두 가지 formulation 으로 표현할 수 있다. Forward diffusion 은 다음과 같다.

$$dx = f(x,t)dt + g(t)dw$$

이는 데이터를 노이즈로 변환하는 과정이며 이 때 x 는 데이터, t 는 diffusion time step, w 는 standard Wiener process, f(x,t)는 drift coefficient, g(t)는 diffusion coefficient 이다. Backward diffusion 은 다음과 같다.

$$dx = [f(x,t) - g(t)^2\nabla_x \log p_t(x)]dt + g(t)d\bar{w}$$

Backward diffusion 은 노이즈를 다시 데이터로 복구하는 과정이며 $\nabla_x \log p_t(x)$ 는 score, 데이터 분포의 gradient 이다. Neural network 를 통해 학습하는 것은 score 이며 이를 통해 backward diffusion process 를 진행하여 노이즈를 데이터로 복구할 수 있다. 실제 score 의 경우

본 논문에서는 제시한 가창 음성 합성 기술은 이 Diffusion 생성 모델을 기반으로 하였으며, text 와 pitch 를 input 으로 받아 그에 해당하는 가창 음성을 생성해 내도록 한다. 제시한 모델의 훈련 과정은 다음과 같다. 먼저 주어진 MIDI file 에서 text 와 pitch 를 추출한다. 그 결과로 text 와 pitch 는 음절 단위로 mapping 된다. 하나의 음절은 초성 중성 종성으로 나뉘어지는데, 이를 초성, 종성에 3 프레임을 할당하고 나머지를 중성에 할당한다. 이는 한국어 발화에서 주로

초성과 종성은 짧은 시간 발음하고 모음에 해당하는 중성에서 대부분의 시간을 할애하는 특징을 이용하였다.

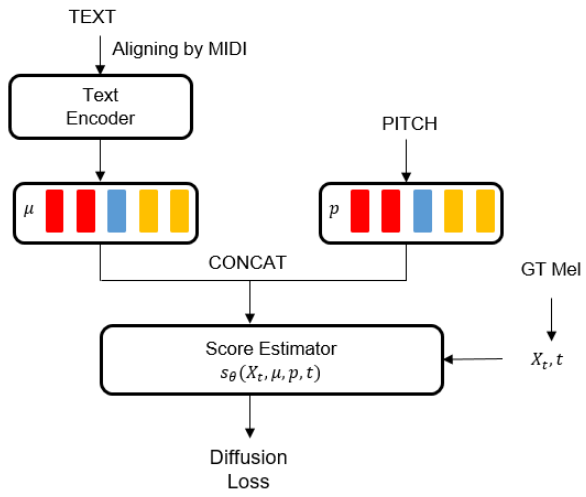

그림 1. 제시한 모델의 훈련 과정

또한 음성 데이터에서 멜-스펙트로그램을 추출하여 이에 임의의 timestep t 에 따른 가우시안 노이즈를 더하여 diffused 된 멜-스펙트로그램을 만들어낸다.
마지막으로 그림 1 에서처럼 이 align 된 text 와 pitch 를 인코더를 통과시킨 뒤에 concatenate 하여 앞서 생성한 diffused 멜-스펙트로그램과 함께 score estimator 에 넣고 실제 score 와 비교하는 diffusion loss 를 줄이도록 학습한다.

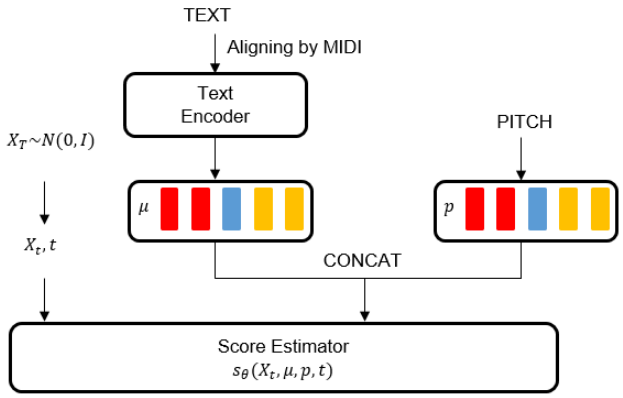

그림 2. 제시한 모델의 생성 과정

생성 과정은 생성하고자 하는 text 와 pitch 를 concatenate 하여 이를 컨디션으로 주고, $N(0, I)$ 에서 샘플링한 노이즈 X_T 에서부터 단계적으로 스코어를 예측하여 결과적으로 멜-스펙트로그램 X_0을 생성한다.

실험은 AI-hub 에서 제공한 다화자 가창 음성 데이터셋을 사용하여 진행되었으며, 베이스라인 모델로는 MLP-singer 을 사용하였다.[1]. MLP-singer 는 MLP-mixer 을 기반으로 한 한국어 가창 음성 합성 모델이다. 그림 3 과 그림 4 는 각각 MLP-singer 와 제시한 모델이 합성한 가창 음성의 멜-스펙트로그램이다. MLP-singer 에 비해 제시된 모델이 생성한 멜-스펙트로그램이 더 비브라토의 표현과 자연스러움 면에서 품질이 우수한 것을 볼 수 있다.

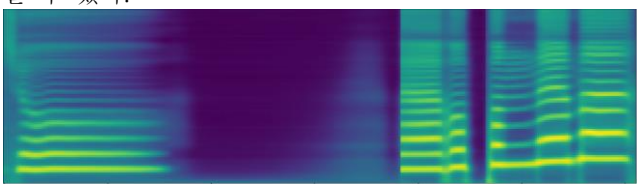

그림 3. MLP-singer 의 멜-스펙트로그램

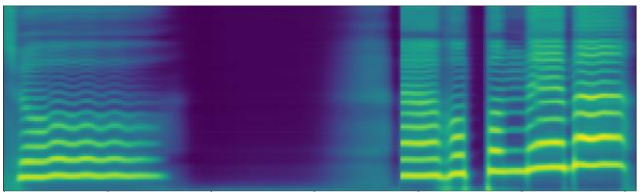

그림 4. 제시한 모델의 멜-스펙트로그램

Ⅲ. 결 론

본 논문에서는 Diffusion 생성 모델을 기반으로 한 한국어 가창 음성 합성 모델을 새로이 제시하여 기존의 모델에 비해 높은 성능을 내는 것을 확인하였다.

ACKNOWLEDGMENT

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

참 고 문 헌

[1] Tae, Jaesung, Hyeongju Kim, and Younggun Lee. "Mlp singer: Towards rapid parallel korean singing voice synthesis." 2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP). IEEE, 2021.

[2] Lee, Juheon, et al. "Adversarially trained end-to-end korean singing voice synthesis system." arXiv preprint arXiv:1908.01919 (2019).

[3] Song, Yang, et al. "Score-based generative modeling through stochastic differential equations." arXiv preprint arXiv:2011.13456 (2020).

[4] Anderson, Brian DO. "Reverse-time diffusion equation models." Stochastic Processes and their Applications 12.3 (1982): 313-326.