디퓨전 생성모델을 활용한 보코더의 샘플링 가속화에 관한 연구

정명훈, 이현승, 김형주, 이동준, 김남수
서울대학교 전기정보공학부 뉴미디어통신공동연구소 휴먼인터페이스 연구실

{mhjeong, hslee, hjkim, djlee}@hi.snu.ac.kr, nkim@snu.ac.kr

A Study on the accelerated sampling of neural vocoder based on diffusion generative model

Myeong Hun Jeong, Hyeon Seung Lee, Hyeong Ju Kim, Dong June Lee, Nam Soo Kim
Human Interface Laboratory,
Department of Electrical and Computer Engineering and INMC,
Seoul National University

요 약

본 논문은 디퓨전 생성 모델을 활용한 보코더의 샘플링 프로세스를 가속화함으로써 모델 성능은 유지하고 샘플링 속도의 향상을 도모하였다. 기존 디퓨전 생성 모델은 양질의 샘플을 얻기 위해서 훈련 과정에서 많은 타임 스텝을 모델이 훈련해야 하는데 이 과정에서 샘플링 속도의 저하가 야기된다. 본 논문에서는 기존의 마르코프 프로세스가 아닌 새로운 디퓨전 프로세스를 제안함으로써 디퓨전 생성 모델 기반 보코더의 샘플링 속도 저하를 개선한다.

Ⅰ. 서 론

본 논문에서는 디퓨전 생성 모델 기반 보코더의 샘플링 속도 개선을 위하여 비 마르코프 디퓨전 프로세스를 새로이 제안한다. 보코더는 음성합성에서 멜 스펙트로그램을 웨이브로 바꿔주는 역할을 한다. 하지만 기존 디퓨전 생성 모델 기반의 보코더는 양질의 샘플을 얻기 위해서 많은 타임 스텝을 모델이 학습해야 하며 이는 곧 샘플링 속도 저하를 야기한다.[1][2] 따라서 본 논문에서는 이러한 디퓨전 생성 모델의 단점을 극복하고 실제 보코더에 사용할 수 있을 만큼 샘플링 속도를 개선하였다.

Ⅱ. 본론

디퓨전 생성 모델은 가우시안 노이즈를 여러 번 더해주는 디퓨전 프로세스와 가우시안 노이즈를 뉴럴 네트워크 모델이 예측하며 제거하는 리버스 프로세스로 이루어져 있다.[3] 이때, 각각의 프로세스는 마르코프 연쇄를 따른다. 디퓨전 프로세스와 리버스 프로세스는 다음과 같다.

$$q(x_1 \ \dots, x_T|\, x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1})$$

$$p_\theta(x_0 \ \dots, x_{T-1}|\, x_T) = \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t), \qquad p_{latent}(x_T) = \mathrm{N}(0, \mathrm{I})$$

디퓨전 생성 모델은 학습과정에서 우도의 로그 값을 최대화하기 위하여 다음과 같은 음의 ELBO(Evidence Lower BOund)를 목적함수로 갖는다.

$$-ELBO = c + \sum_{t=1}^{T} k_t E_{x_0,\epsilon}\left\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\right\|_2^2$$

디퓨전 생성 모델의 샘플링 과정은 모델이 예측한 가우시안 노이즈를 점차적으로 제거하면서 x0 방향으로 진행하게 된다. 샘플링 식은 다음과 같다.

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right) + \sigma_t z$$

이렇게 샘플링을 하게 되면 최종적인 x0 를 얻기 위해서 기 설정한 타입스텝을 모두 거쳐서 샘플링을 하게 된다. 디퓨전 생성 모델은 기본적으로 많은 타임 스텝을 모델이 학습할수록 더 좋은 성능을 내게 되는데, 좋은 성능을 내기 위해서 샘플링 과정이 느려질 수밖에 없다는 단점이 있다. 따라서 아래와 같은 비 마르코프 연쇄에 따른 디퓨전 프로세스를 뉴럴 보코더 모델에 새롭게 제안한다.[4]

$$q_\sigma(x_1 \ \dots, x_T|\, x_0) := q_\sigma(x_T|x_0)\prod_{t=2}^{T} q_\sigma(x_{t-1}|x_t, x_0)$$

위 식처럼 디퓨전 프로세스를 가정하게 되면, 아래와 같은 식으로 목적식이 유도된다.

$$J_\sigma(\epsilon_\theta) \equiv \sum_{t=1}^{T}\frac{1}{2d\sigma_t^2\alpha_t}E\left[\left\|\epsilon_\theta^{(t)}(x_t) - \epsilon_t\right\|_2^2\right]$$

위 목적식은 원래 디퓨전 기반의 목적식과 유사한 것을 확인할 수 있다. 따라서 비 마르코프 연쇄로 가정하더라도 기존에 훈련한 모델을 그대로 사용할 수 있다. 본 논문은 이러한 사실에 착안하여 똑같은 목적식을 갖는 새로운 샘플링 방법을 디퓨전 생성모델 기반 뉴럴 보코더에 적용한 것이 주요 아이디어이다. 새롭게 제안한 샘플링 과정은 다음과 같이 나타낼 수 있다.

$$x_{\tau-1} = \sqrt{\alpha_{\tau-1}}\left(\frac{x_\tau - \sqrt{1-\alpha_\tau}\epsilon_\theta^{(\tau)}(x_\tau)}{\sqrt{\alpha_\tau}}\right) + \sqrt{1-\alpha_{\tau-1}-\sigma_\tau^2}\epsilon_\theta^{(\tau)}(x_\tau) + \sigma_\tau\epsilon_\tau$$

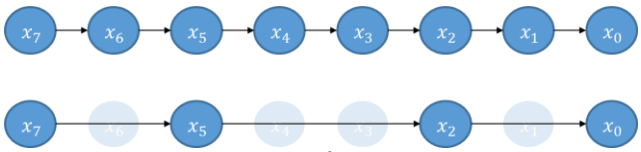
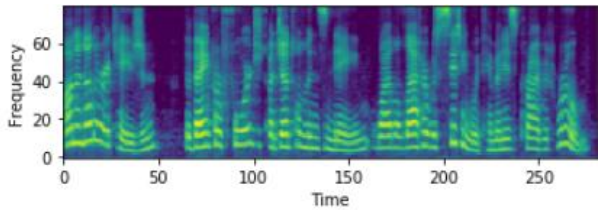


그림 1

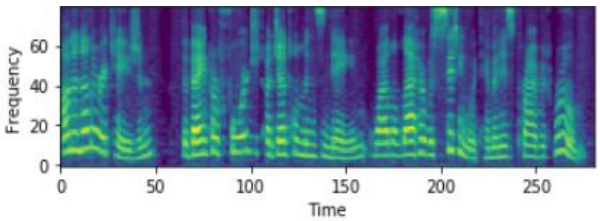
중요한 사실은 새로운 샘플링 식에서, 원래의 타임 스텝 전체가 아닌, subsequence 에 대해 샘플링을 진행하더라도 비슷한 샘플 퀄리티를 얻게 된다는 것이다. 대략적인 샘플링 과정은 그림 1 에 나타나 있다. 그림 1 에서는 7 타임 스텝으로 훈련하고 4 개의 subsequence 로 샘플링 한 것을 도식화 한 것이다. 본 논문에서는 200 타임 스텝으로 훈련하고 10 타임 스텝씩 건너뛰며 샘플링 하여 총 20 타임 스텝으로 샘플링을 진행하였다. 24 시간 LJSpeech Dataset 을 활용하였고, [1]에서 사용한 보코더 모델을 베이스라인으로 사용하여 실험을 진행하였다. 1 초의 음성을 생성하는데 걸리는 시간(sec)을 RTF(Real Time Factor)로 설정하여 표 1 과 같은 결과를 얻었다. 기존의 디퓨전 생성모델 기반의 보코더는 200 타임스텝으로 샘플링 할 때, 5.95 의 RTF 로 매우 느린 것을 확인할 수 있는데, 본 논문에서 제안한 샘플링 가속화 방법을 사용하면 0.53 의 RTF 로 T=50 일 때의 샘플링 보다 오히려 더 빠른 샘플링 속도를 확인할 수 있었다.
또한 그림 2 에서 알 수 있듯이, 200 타임 스텝으로 훈련하고 200 타임 스텝으로 샘플링 한 것과 제안한 방법으로 샘플링 한 것을 비교했을 때 샘플의 질은 거의 차이가 없는 것을 확인할 수 있었다.

Model	RTF
DDPM(T=5)	0.05
DDPM(T=50)	0.72
DDPM(T=200)	5.95
Proposed	**0.53**

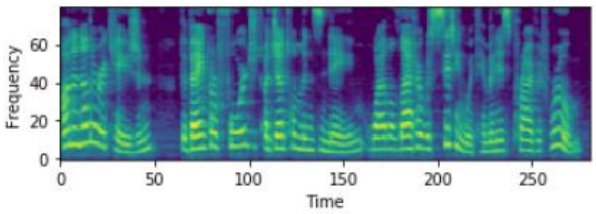
표 1 샘플링 방법에 따른 RTF



(a) T=200 으로 샘플링한 결과



(b) 제안한 방법으로 샘플링한 결과



(c) 원본

그림 2 T=200 으로 훈련한 모델로 샘플링 방식을 다르게 생성한 음성의 로그 멜스펙트로그램

Ⅲ. 결론

본논문에서는 비 마르코프 연쇄 가정을 함으로써 디퓨전 생성모델 기반 보코더의 샘플링 속도를 개선하고 모델의 성능을 유지하는 새로운 샘플링 방법을 제안하였다. 실험을 통해 제안된 샘플링 방법을 적용했을 때 더 빠른 샘플링 속도를 확인할 수 있었고, 모델의 성능도 유지할 수 있다는 것을 보였다. 이를 통해 디퓨전 생성모델을 기반으로 한 보코더도 속도에 제한없이 양질의 샘플을 얻을 수 있게 되었다는 데 의의가 있다.

ACKNOWLEDGMENT

이 논문은 2021 년도 BK21 FOUR 정보기술 미래인재 교육연구단에 의하여 지원되었음.

참 고 문 헌

[1] Kong, Zhifeng, et al. "DiffWave: A Versatile Diffusion Model for Audio Synthesis." arXiv preprint arXiv:2009.09761 (2020).

[2] Chen, Nanxin, et al. "WaveGrad: Estimating gradients for waveform generation." arXiv preprint arXiv:2009.00713 (2020).

[3] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." Advances in Neural Information Processing Systems 33 (2020).

[4] Song, Jiaming, Chenlin Meng, and Stefano Ermon. "Denoising Diffusion Implicit Models." arXiv preprint arXiv:2010.02502 (2020).