잡음 레벨 정렬을 활용한 디퓨전 음성 향상

전용현, 우범준. 김정훈, 김남수
서울대학교
{yhjeon, bjwoo, jhkim}@hi.snu.ac.kr, nkim@snu.ac.kr

Aligning Noise Level For Diffusion Speech Enhancement

Yong Hyeon Jun, Beom Jun Woo, Jeunghun Kim, and Nam Soo Kim
Department of Electrical and Computer Engineering and INMC, Seoul National Univ.

요 약

Ⅰ. 서 론

최근 음성 향상 기술은 다양한 환경에서 음성 품질을 향상시키는 데 중점을 두고 발전해왔다. 전통적인 노이즈 억제 및 에코 제거 기법은 음성 신호의 왜곡 문제를 해결하지 못하는 한계를 지니고 있다. 이러한 한계를 극복하기 위해 생성형 모델 기반의 디퓨전 모델[1]이 음성 향상 분야에서 주목받고 있다[2]. 특히 디퓨전 모델은 데이터를 점진적으로 변환하며 잡음을 제거하는 데 효과적이다. 본 연구는 디퓨전 모델의 성능을 향상시키기 위해 잡음 레벨 정렬 기법을 제안한다. 이는 환경 잡음과 가우스 잡음 간의 비율을 일정하게 유지하여 잡음 제거 성능을 최적화한다. 이를 통해 음성의 왜곡을 최소화하고 청취 품질을 크게 향상시킬 수 있다.

Ⅱ. 본론

1) 잡음 레벨 정렬 디퓨전과 그 역방향 과정

본 연구는 잡음 레벨 정렬을 통해 전방 과정에서 환경 잡음과 가우스 잡음의 비율을 일정하게 유지하는 새로운 확률 미분방정식을 제안한다. 전방 과정 $dx_t = f(t)dt + g(t)dw$로 표현할 때, $f(t) = \gamma(y - x_t)$인 경우 $g(t) = \sigma_{\max}\sqrt{2\gamma(1-e^{-\gamma t})}$로 계산되어 $dx_t = \gamma(y - x_t)dt + \sigma_{\max}\sqrt{2\gamma(1-e^{-\gamma t})}dw$로 표현되며, $f(t)$가 변화함에 따라 그에 맞는 $g(t)$를 계산하여 다양한 $f(t)$에 대하여 유동적으로 적용할 수 있다.

이 방정식은 시간에 따라 환경 잡음과 가우스 잡음이 일정한 비율로 증가하도록 보장한다. 전방 과정은 음성 신호의 명료성을 유지하면서 잡음을 점진적으로 증가시키는 방식으로 설계된다. 이러한 설정은 잡음 제거 과정에서 발생할 수 있는 왜곡을 최소화하고, 음성 신호의 자연스러움을 유지하는 데 기여한다.

이 과정에서 디퓨전 계수는 시간에 따라 변화하는 가우스 잡음의 분산 경로를 제어한다. 초기 단계에서는 빠르게 증가하여 잡음을 많이 추가하고, 후반 단계에서는 점차 완화되어 안정적인 신호를 유지한다. 이는 기존 연구에서 제안된 단순한 디퓨전 스케줄과의 차별점이다.

역방향 과정은 제안된 전방 과정에 따라 설정되며, 이를 통해 초기 잡음 데이터에서 깨끗한 음성 데이터를 복원한다. 역방향 과정은 다음과 같이 정의된다:

$$dx_t = [f(x_t,t) - g^2(x_t,t)s_\theta(x_t,t)]dt + g(x_t,t)dw_t$$

이 과정은 디퓨전 모델의 효율성을 극대화하고 복원된 음성 신호의 품질을 높이는 데 중점을 둔다. 환경 잡음과 가우스 잡음 간의 균형을 유지하는 점에서 기존 방식보다 안정적이다. 특히, 전방 과정에서 설정된 균형을 역방향 과정에서도 유지함으로써, 불필요한 음성 신호의 왜곡을 줄일 수 있다.

역방향 과정에서 사용되는 스코어 함수는 모델이 학습을 통해 예측한 신호의 기울기를 나타낸다. 이는 각 시간 단계에서 복원해야 할 신호의 방향을 제공하며, 디퓨전 모델의 성능에 중요한 영향을 미친다.

2) 모델 학습 및 샘플링

디퓨전 모델의 학습 과정은 제안된 전방 과정과 역방향 과정을 기반으로 한다. 학습 단계에서는 스코어 모델이 환경 잡음과 가우스 잡음 간의 비율을 유지하도록 설계된다. 이 모델은 다음 손실 함수를 최소화하도록 학습된다:

$$L = E_{t,x_t}[\parallel \nabla x_t \log p_t(x_t) - s_\theta(x_t,t) \parallel^2]$$

이 손실 함수는 학습 과정에서 스코어 모델이 잡음이 포함된 신호에서 깨끗한 신호로의 경로를 효과적으로 학습하도록 돕는다. 특히, 이상적인 복원 경로를 나타내며, 모델이 이를 학습하도록 하는 것이 핵심이다.

샘플링 단계에서는 학습된 모델을 사용하여 역방향 과정을 수행하며, 이를 통해 깨끗한 음성을 복원한다. 샘플링 과정은 효율적인 계산을 통해 실시간 응용에도 적합하다. 특히, 제안된 방법은 디퓨전 과정의 반복 횟수를 줄이면서도 높은 품질의 결과를 제공한다.

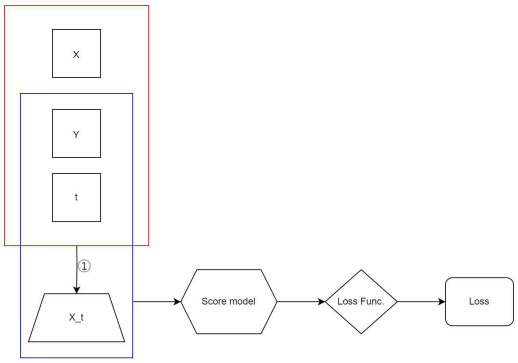

그림 1. 디퓨전 음성 향상 모델 학습 과정

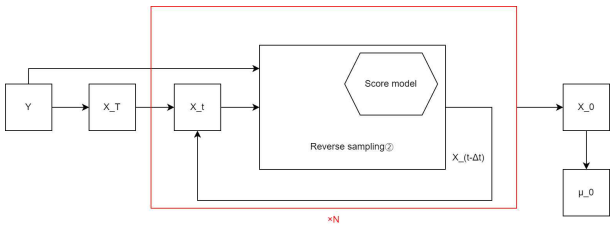

그림 2 디퓨전 음성 향상 수행 과정

3) 실험 및 결과

본 연구는 다양한 잡음 환경에서 제안된 방법의 성능을 검증하였다. 학습은 Voicebank-DEMAND[3] 학습 데이터셋을 기반으로 했으며, 동일한 데이터의 검증 데이터셋에 대해서 성능 평가를 진행했다. 성능 평가 지표로는 PESQ[4](Perceptual Evaluation of Speech Quality)를 사용하였다. 실험 환경은 일반적인 통신 환경과 강한 잡음 조건을 모두 포함하여 설정되었다. 특히, 다양한 유형의 잡음을 추가하여 제안된 기법의 일반화 성능을 평가하였다.

실험 결과, 제안된 방법은 역방향 과정을 충분히(N=30) 반복했을 시 기존 디퓨전 기반 음성 향상 모델[2]과 확연한 성능 차이가 없었지만, 연산량 감소를 위해 역방향 과정 횟수를 기존의 1/2로 축소한 경우(N=15) PESQ 성능 평가에서 큰 향상을 보였다. 이는 역방향 과정을 줄여 경량화된 디퓨전 모델에 있어서 잡음 정렬 기법이 성능을 개선했음을 나타낸다. 검증 데이터셋은 학습 데이터의 잡음 환경과는 다른 잡음 환경이 포함되어있어, 제안된 모델은 학습 데이터의 잡음 환경과 다른 새로운 잡음 환경에서도 강인한 성능을 유지하는 것을 확인할 수 있었다.

	기존 음성 향상 모델	제안된 모델
PESQ(N=30)	3.04	3.02
PESQ(N=15)	2.07	**2.73**

표 1 성능 평가 결과

본 연구에서 제안된 기법은 단순히 음성 향상에 국한되지 않는다. 예를 들어, 음성 기반 인공지능 비서 시스템, 화상회의 소프트웨어, 차량 내 음성 인식 시스템 등에서도 잡음 제거 기술로 활용 가능하다. 또한, 제안된 모델은 디퓨전 과정의 계산 효율성을 유지하면서도 높은 성능을 제공하기 때문에 저자원 환경에서도 활용이 가능하다. 이러한 응용은 디퓨전 모델의 실질적인 가치를 높이며, 다양한 산업 분야에서 유용한 도구로 자리잡을 수 있음을 시사한다.

Ⅲ. 결론

본 연구는 디퓨전 음성 향상 모델의 성능을 극대화하기 위해 잡음 레벨 정렬 기법을 제안하였다. 제안된 방법은 환경 잡음과 가우스 잡음 간의 비율을 일정하게 유지함으로써 음성 신호의 품질을 개선하였다. 특히, 제안된 기법은 디퓨전 과정에서 발생할 수 있는 비효율성을 줄이고 실시간 응용 가능성을 높였다. 잡음 정렬 기법은 음성 향상 외에도 뇌파 데이터, 이미지 및 비디오 데이터 등 다양한 응용 가능성을 가지고 있다. 이러한 연구는 디퓨전 모델의 응용 범위를 넓히고, 다양한 분야에서 기술적 혁신을 가져올 것으로 기대된다.

ACKNOWLEDGMENT

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

참 고 문 헌

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

[2] Welker, Simon, Julius Richter, and Timo Gerkmann. "Speech enhancement with score-based generative models in the complex STFT domain." arXiv preprint arXiv:2203.17004 (2022).

[3] Valentini-Botinhao, Cassia, et al. "Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech." SSW. 2016.

[4] Rix, Antony W., et al. "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs." 2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221). Vol. 2. IEEE, 2001.