Wavenet을 이용한 음성 신호의 대역폭 확장

*서은미 유정찬 임유진 박호종
광운대학교
*dlvbfl0513@gmail.com

Speech Bandwidth Extension using Wavenet

*Seo, Eunmi Yu, Jeongchan Lim, Yujin Park, Hochong
Kwangwoon University

요약

본 논문은 wavenet을 이용하여 음성 신호의 대역폭을 확장하는 새로운 모델을 제안한다. 기존의 대역폭 확장은 주로 주파수 영역에서 진행되며, 확장 대역의 주파수 크기는 높은 성능으로 복원하지만 위상 정보를 정확히 복원할 수 없다는 문제점을 가진다. 이를 해결하기 위해 wavenet 모델을 기반으로 시간 영역에서 저대역과 고대역의 상관관계를 이용하여 고대역 성분을 생성하도록 한다. 제안하는 방법은 모든 동작을 시간 영역에서 수행하며, 제안 방법으로 생성한 고대역 성분이 원음의 고대역 성분과 유사한 것을 확인하였다.

1. 서론

현재 오디오 신호처리 분야에서 인공지능을 활용한 연구가 증가하고 있다. 그 중 최근 발표된 wavenet은 기존의 모델과는 달리 dilated convolution을 사용하여 시간 영역에서 엔드-투-엔드 (end-to-end)로 동작하며 좋은 성능을 보이고 있다[1]. 따라서 음성과 관련하여 다양한 분야에서 이를 활용한 다양한 연구가 진행되고 있다.

샘플링 주파수를 높이기 위해 사용되는 기존 신호처리 방법으로는 싱크 (sinc) 보간법이 있으나, 고대역 성분은 생성할 수 없다는 단점이 존재한다. 이에 따라 주파수 영역에서 다양한 처리 방법을 사용하여 대역폭을 확장하는 방법이 개발되었다. 하지만 주파수 영역의 처리에서는 위상 정보와 관련하여 다양한 문제점들이 제기되었다.

따라서 본 논문은 wavenet 모델을 이용하여 시간 영역에서 업 샘플링 (up-sampling)과 대역폭 확장하는 방법을 제안한다. 이는 시간 영역에서 end-to-end로 동작하므로 위상 정보가 손실되지 않도록 한다. 또한 제안하는 모델을 통해 생성된 음원의 고대역 성분이 동일한 샘플링 주파수를 가지는 원본 음원의 고대역 성분과 비슷한 것을 스펙트로그램을 통해 확인하였다.

2. 제안하는 방법

2.1 Denoising Wavenet

제안하는 wavenet 모델은 denoising wavenet 구조를 기반으로 한다[2]. 한 샘플씩 생성하는 일반적인 wavenet 모델과는 다르게 denoising wavenet 모델은 프레임 단위로 처리가 가능하다.

일반적인 wavenet 모델은 과거의 샘플만을 이용하는 인과 (causal) dilated convolution을 통해 새로운 샘플을 생성한다. 이와 달리 denoising wavenet 모델은 과거 샘플과 미래 샘플을 모두 이용하는 비인과 (non-causal) dilated convolution을 사용한다. Denoising wavenet에서 사용되는 dilated convolution의 예시는 그림 1과 같다.

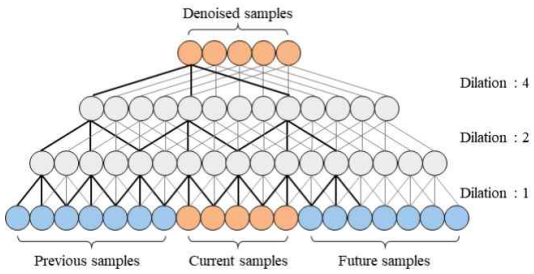

그림 1. Denoising wavenet에서의 dilated convolution 예시
Fig. 1. Example of dilated convolution in denoising wavenet

Denoising wavenet의 전체 구조는 그림 2와 같다. 이 때 각 dilated convolution 층은 residual block으로 이루어지며, dilated convolution을 이용하는 gated unit과 residual, skip connection을 위한 convolutional neural network (CNN) 단이 포함된다.

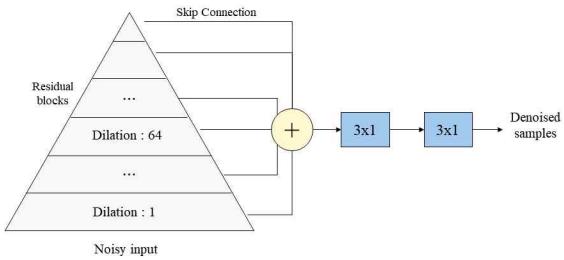

그림 2. Denoising wavenet의 구조도
Fig. 2. Structure of denoising wavenet

2.2 제안하는 모델의 구조

제안하는 모델의 입력은 싱크 보간법을 통해 원하는 샘플링 주파수를 가지는 음원으로 변경된다. 변경된 음원은 네트워크의 입력으로 사용되며 출력으로는 대역폭 확장된 음원을 얻는다.

제안하는 모델의 구조도는 그림 3과 같다. 싱크 보간법, gated unit, residual, skip connection과 입출력에서의 CNN 단으로 구성되어있다. Residual의 경우 각 단의 출력이 입력과 더해져 다음 단의 입력으로 사용될 수 있도록 한다. Skip connection은 각 단의 출력이 가지는 특성들을 종합하여 최종 예측에 도움이 되도록 하며, 단이 깊어짐에 따라 발생할 수 있는 기울기 소실 (gradient vanishing)을 방지한다.

제안하는 모델에서는 과적합이 발생하지 않도록 하기 위해 레이어 정규화 (layer normalization)를 추가한다[3]. 레이어 정규화는 gated unit 이후와 skip out 이후로 총 2번 진행하였다.

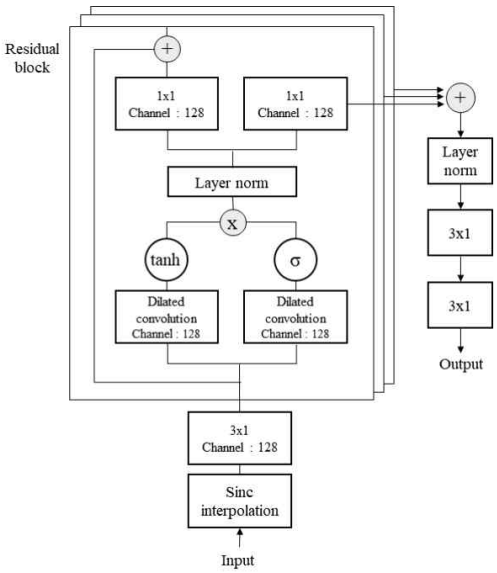

그림 3. 제안하는 모델의 구조도

Fig. 3. Structure of proposed model

3. 성능 평가

Wavenet 학습과 성능 평가는 화자 1명에 대한 약 36분 길이의 한국어 음성 데이터를 이용하여 진행하였다. 입력의 샘플링 주파수는 8 kHz, 출력의 샘플링 주파수는 16 kHz로 샘플링 주파수를 2배 증가시켰다.

모델은 총 20개의 dilated 단으로 구성되어 있다. 각 단의 dilation은 1에서 512까지 2의 제곱수로 증가하며, 이를 2번 반복한다. 수용 필드 (receptive field)는 6,139개의 샘플, 타겟 필드 (target field)는 1,601개의 샘플로 구성된다. 따라서 약 384ms의 음원을 이용하여 약 100ms의 음원을 생성한다. 손실함수는 mean absolute error (MAE), 최적화는 Adam을 사용하여 진행하였다.

그림 4는 원본과 싱크 보간법을 이용하여 8 kHz에서 16 kHz로 업 샘플링 된 음원, 제안하는 모델에 대한 결과의 스펙트로그램이다. 싱크 보간법을 이용하여 업 샘플링 된 음원의 경우, 고대역 성분이 존재하지 않는 것을 볼 수 있다. 반면 업 샘플링 이후 모델을 통과한 음원에서는 원본 음원과 유사한 형태의 고대역 성분이 생성되었다. 이를 통해 해당 모델이 시간 영역에서 저대역과 고대역의 상관관계를 분석하고 입력에 대한 대역폭 확장을 수행하는 것을 확인하였다.

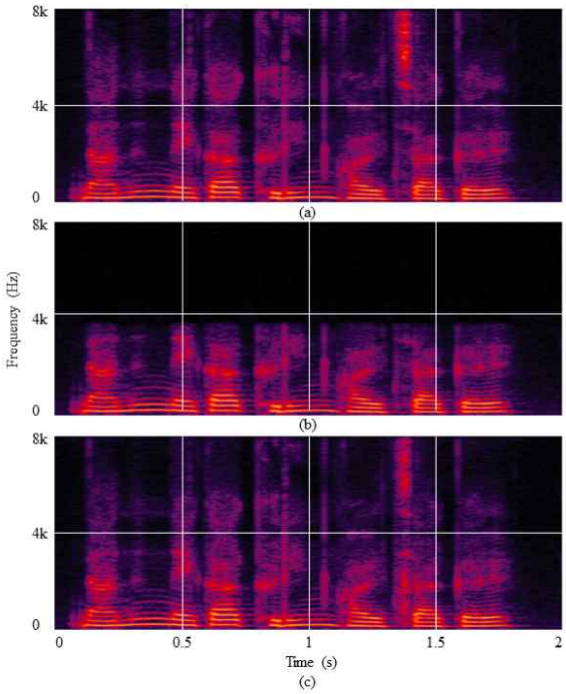

그림 4. 스펙트로그램 (a) 16 kHz 원본 음원 (b) 싱크 보간법을 이용하여 업 샘플링 된 음원 (c) 제안하는 모델을 통해 대역폭 확장된 음원

Fig. 4. Spectrogram of (a) 16 kHz original speech (b) Up-sampled speech using sinc interpolation (c) Speech with extended bandwidth using proposed model

4. 결론

본 논문은 wavenet 모델을 이용하여 대역폭을 확장하는 방법을 제안하였다. 네트워크의 입력으로는 업 샘플링 된 신호가 사용되며, 과적합을 막기 위해 레이어 정규화가 사용된다. 제안하는 모델을 통해 생성된 음원과 원본 음원의 고대역 성분이 유사함을 스펙트로그램을 통해 확인하였다.

감사의 글

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No. 2017-0-00072).

참고문헌

[1] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. “Wavenet: A generative model for raw audio,” *arXiv:1609.03499*, 2016.

[2] D. Rethage, J. Pons, and X. Serra, “A Wavenet for Speech Denoising,” *arXiv:1706.07162*, 2018.

[3] J. Lei Ba, J. Ryan Kiros, Geoffrey E. Hinton, “Layer Normalization,” *arXiv:1607.06450*, 2016.