

## 조건부 Wavenet을 이용한 음성 신호의 잡음 제거 기술

\*유정찬 서은미 임유진 박호종

광운대학교

\*yjc133@naver.com

## Speech Denoising using Conditional Wavenet

\*Yu, Jeongchan Seo, Eunmi Lim, Yujin Park, Hochong

Kwangwoon University

## 요약

본 논문에서는 조건부 wavenet을 이용한 음성 신호의 잡음 제거 기술을 제안한다. 기존의 음성 신호 잡음 제거 기술은 스펙트로그램을 기반으로 발전되어 왔으나, 잡음으로 인해 변형된 원음의 위상 정보를 복원할 수 없는 문제점을 가진다. 이를 해결하기 위해 시간 영역에서 전 과정을 실행하는 기계학습 모델인 wavenet을 사용하여 음성 신호의 잡음을 제거하는 방법을 제안한다. 특히, 잡음 종류를 조건으로 입력하여 성능 향상을 얻도록 한다. 성능 평가를 통하여 제안 방법이 시간 영역에서 잡음을 감소시킬 수 있음을 확인하였다.

## 1. 서론

인공지능의 발달로 오디오 신호처리 기술에 인공지능을 적용한 다양한 연구가 진행되고 있다. 특히 음성 합성, 음성 인식과 같은 음성 신호 처리 분야에서 좋은 성능을 보이며 많은 주목을 받고 있다[1].

기존의 인공지능 기반 음성 신호처리 기술은 대부분 스펙트로그램을 기반으로 하는 프런트 엔드(front-end) 방식으로 진행되었다. 이러한 방식은 스펙트로그램의 증폭 정도는 조절 할 수 있지만 위상 정보를 조절할 수 없다는 문제를 가진다[2].

최근 발표된 기계학습 모델인 wavenet은 dilated convolution을 사용하며 뛰어난 성능으로 시간 영역에서 종단 간(end-to-end)으로 음성 신호를 생성할 수 있음을 보였다. 또한 음성 인식, 음성 변조와 같이 음성과 관련된 분야에서도 좋은 성능을 보여 이를 활용한 많은 연구가 활발히 진행되고 있다[2-3].

본 논문에서는 최근 음성 신호처리에서 좋은 성능을 보이는 wavenet 모델을 이용하여 음성 신호의 잡음을 제거하고자 한다. 시간 영역에서 end-to-end로 동작하여 원본 음성의 위상 정보 손실을 최소화 하고자 하였으며, 결과물에서 잡음이 감소하는 것을 확인하였다. 또한 조건부로 잡음의 종류를 입력하여 성능을 향상시킬 수 있음을 확인하였다.

## 2. 제안하는 방법

## 2.1 제안하는 모델의 구조

제안하는 모델은 wavenet 모델을 기반으로 한다. Wavenet 모델은 dilated convolution을 사용하여 한 샘플을 생성할 때 매우 넓은 범위의 시간 축 신호를 비교적 적은 연산량으로 컨볼루션 할 수 있다는 장점을 가진다.

일반적인 wavenet 모델에서는 과거의 값만 사용하는 인과

(causal) dilated convolution을 통해 음성 신호를 생성한다. 하지만 제안하는 모델에서는 비인과(non-causal) dilated convolution을 사용함으로써 과거와 미래의 값을 이용하여 음성 신호를 생성한다. 따라서 커널 크기가 2인 기존의 wavenet 모델과 달리 제안하는 모델은 그림 1과 같이 커널 크기로 3을 사용한다.

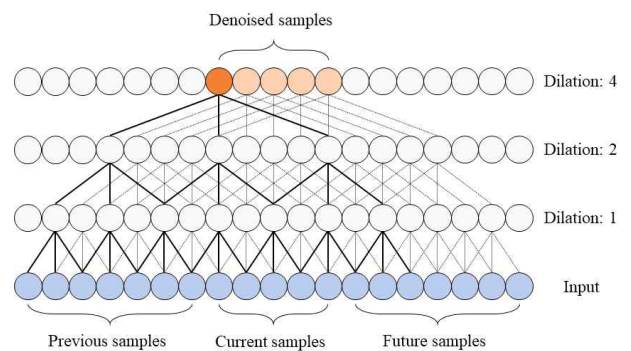


그림 1. 제안하는 모델에서의 dilated convolution 구조

Fig. 1. Structure of dilated convolution in the proposed model

Wavenet 모델은 dilated convolution 외에도 gated unit과 residual, skip connection을 포함하고 있다. 각각의 층은 residual block이라 부르며 dilated convolution이 적용된 gated unit과 residual, skip connection을 포함한다.

제안하는 모델 또한 이러한 구조를 차용하였다. 그림 2는 제안하는 모델의 전체 구조이며, 그림 3은 residual block의 구조이다.

## 2.2 조건부 적용 방법

제안하는 방법에서는 잡음 종류를 one-hot 벡터로 변환 후 convolutional neural network (CNN)으로 차원을 확장하여 gated unit의 입력에 더함으로써 조건을 추가할 수 있도록 하였으며 식 (1)과

같이 표현한다. 이때 잡음 종류를 one-hot으로 나타낸 벡터를  $h$ ,  $h$ 의 차원을 확장하도록 학습된 CNN을  $f$ , gated unit의 입력을  $x$ , 학습된 dilated convolution의 가중치를  $W$ , gated unit의 출력을  $z$ 로 나타낸다.

$$z = \tanh(W_t^* x + f_t(h)) \odot \sigma(W_s^* x + f_s(h)) \quad (1)$$

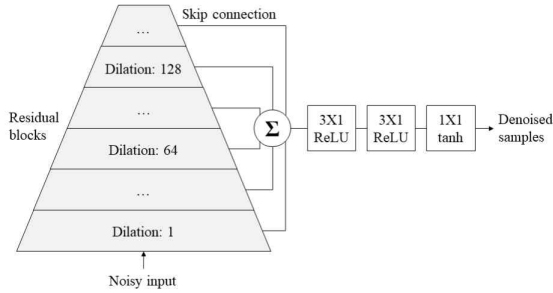


그림 2. 제안하는 모델의 전체 구조도

Fig. 2. Overall structure of the proposed model

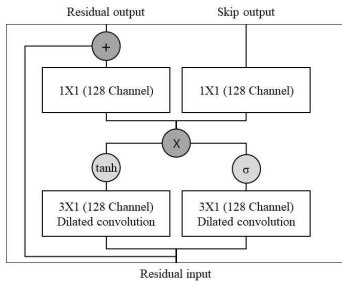


그림 3. Residual block의 구조도

Fig. 3. Structure of residual block

### 3. 성능 평가

성능 평가를 위해 최종적으로 출력된 신호의 SNR, segmental SNR (SSNR), PESQ를 측정하였다. 학습과 성능 평가를 위한 신호는 16 kHz로 샘플링 된 화자 1 명의 한국어 음성 신호이고, 원음에 잡음을 0 dB SNR로 혼합하여 사용하였다. 잡음은 babble, car, metro, white로 총 4 가지를 사용하였다. 데이터 크기는 총 18분이며 학습에 15분, 성능 평가에 3분을 사용하였다.

Wavenet 모델에서 receptive field의 크기는 6000 샘플이며, 그 중 1600 샘플의 잡음을 제거한다. 은닉 층의 dilation은 [1, 2, 4, 8, 16, 32, 64, 128, 256, 512]로 2의 제곱수로 증가시켰으며 이를 2 회 반복 하였다. 모델의 학습은 mean absolute error (MAE)를 손실 함수로 하여 Adam을 사용하여 진행하였다.

표 1은 잡음 제거 전과 제안하는 방법으로 잡음을 제거한 후의 SNR, SSNR, PESQ를 나타낸다. 제안하는 방법을 사용하였을 때, 입력 된 신호에 비해 평균 11.0 dB의 SNR, 13.4 dB의 SSNR 증가를 보였고, 조건부를 추가할 경우 입력 신호에 비해 평균 13.1 dB의 SNR, 15.7 dB의 SSNR의 증가로 성능이 더 향상되는 것을 확인하였다.

그림 4는 잡음이 제거되기 전과 제안 방법을 통해 잡음이 제거된 후의 스펙트로그램 예시이다.

표 1. 모델별 잡음 감소 성능 비교

Table 1. Comparison of noise reduction performance for each model

Noise Type	Evaluation Model	SNR (dB)	SSNR (dB)	PESQ
Babble	None	0.0	-8.0	1.7
	Wavenet	8.0	1.5	2.1
	<b>Wavenet_cond</b>	<b>10.1</b>	<b>4.3</b>	<b>2.2</b>
Car	None	0.0	-5.7	3.8
	Wavenet	12.3	9.4	2.2
	<b>Wavenet_cond</b>	<b>15.9</b>	<b>12.3</b>	<b>3.0</b>
Metro	None	0.0	-4.7	2.2
	Wavenet	11.8	8.4	2.2
	<b>Wavenet_cond</b>	<b>14.2</b>	<b>10.9</b>	<b>2.7</b>
White	None	0.0	-8.4	1.29
	Wavenet	11.8	7.8	2.1
	<b>Wavenet_cond</b>	<b>12.0</b>	<b>8.5</b>	<b>2.3</b>

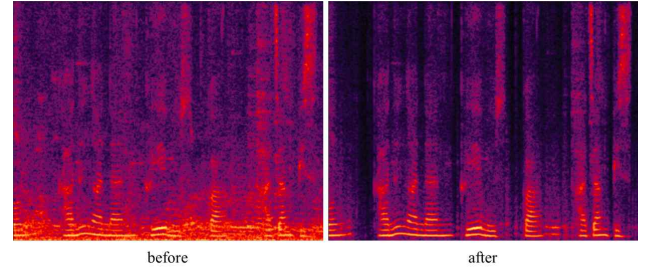


그림 4. 잡음 제거 전과 후의 스펙트로그램

Fig. 4. Spectrogram before and after denoising

### 4. 결론

본 논문에서는 조건부 wavenet 모델을 이용하여 음성 신호의 잡음을 제거하는 방법을 제안하였다. 시간 영역에서 end-to-end로 동작하며 잡음이 감소된 음성 신호를 생성하는 것을 확인하였고, 조건부로 잡음 종류를 입력하여 잡음 제거 성능을 향상시킬 수 있음을 확인하였다.

### 감사의 글

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

### 참고문헌

- [1] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.Y. Chang, and T. Sainath, "Deep Learning for Audio Signal Processing," *IEEE J. of Selected Topics in Signal Processing*, Vol. 13, No. 2, pp. 206-219, May 2019.
- [2] D. Rethage, J. Pons, and X. Serra, "A Wavenet for Speech Denoising," *arXiv:1706.07162*, 2018.
- [3] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and Ko. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv:1609.03499*, 2016.