논문 19-44-05-10 The Journal of Korean Institute of Communications and Information Sciences '19-05 Vol.44 No.05
https://doi.org/10.7840/kics.2019.44.5.864

잡음제거 모델 훈련을 위한 딥러닝 기반 가상 데이터베이스 생성 기법

윤 덕 규*, 최 승 호°

Deep Learning-Based Virtual Database Creation Techniques for Denoising Model Training

Deokgyu Yun*, Seung Ho Choi°

요 약

딥러닝 기반 잡음제거를 위한 신경망 훈련에는 실제 잡음환경에서 취득한 대량의 데이터가 필요하지만 비용 등 여러 측면에서 용이하지 않다. 본 연구는 이에 대한 대안 방법으로서 우선, 실제 잡음환경에서 적당한 양의 원음 및 혼합신호 데이터를 이용하여 심층신경망 기반으로 환경을 추정한다. 이후 추정된 환경 즉, 잡음환경 심층신경망을 이용하여 원음 스펙트럼을 입력으로 스펙트럼 비율(ideal ratio mask)을 출력으로 하는 방법으로 가상의 혼합신호 데이터를 생성한다. 제안한 방법을 통해 실제 환경과 유사한 대량의 데이터베이스를 구축할 수 있으며, 이를 통해 잡음제거 모델의 성능을 크게 개선할 수 있다. 실제 환경에서의 실험을 통해, 딥러닝 기반 잡음제거를 위해 제안한 방법이 성공적으로 사용될 수 있음을 보였다.

Key Words : Deep neural network, virtual noisy database, real environment, denoising, ideal ratio mask

ABSTRACT

Neural network training for deep learning-based noise cancellation requires a large amount of data acquired in a real noise environment, but it is not easy in many respects, such as various costs. In this paper, we propose an alternative method to estimate the environment based on the deep neural network using the proper amount of original sound and noisy signal data. Then, it is possible to generate a large amount of virtual noisy database by inputting the original sound spectrum through the deep neural network and outputting the ideal ratio mask. We can build a large database similar to the actual environment through the proposed method, which can greatly improve the denoising performance. Experiments in real environments have shown that the proposed method can be used successfully for deep learning-based denoising.

Ⅰ. 서 론

잡음환경에서의 취득한 음성신호에서 잡음을 제거하기 위해 심층신경망을 활용하는 방법은 주로 잡음이 혼합된 신호의 스펙트럼을 입력으로 하여 원음의 스펙트럼이 출력되도록 훈련시키거나[1] 스펙트럼의 비율 (ideal ratio mask, IRM)[2]을 출력하도록 훈련시킨다. 이와 같은 방식의 경우, 심층신경망 훈련에 사용하는 데이터는 주변 소음, 잔향 환경, 마이크 특성 등이 실제 사용 환경과 서로 유사한 환경에서 취득한 것일수록 잡음제거 성능이 높다. 따라서 인위적으로 원음과 잡음을 더한 시뮬레이션 데이터가 아닌 실제 환경에서 취득한 데이터베이스를 사용하여 잡음제거 모델을 훈련할 필요가 있다[3]. 그러나 실제 환경에서 대량의 데이터를 취득하는 것은 많은 시간과 비용이 소모되며, 다양한 종류의 음성신호를 취득하는데 어려움이 있다.

본 논문에서는 원음으로부터 실제 환경과 유사한 가상의 혼합신호 데이터베이스를 구성하여 잡음제거 모델을 훈련시키는 방법을 제안한다. 이를 위해 데이터베이스를 구축하는 방법은 실제 잡음 환경에서 재생중인 신호를 취득하여 혼합 신호로 사용하며, 이렇게 취득한 혼합 신호 중에서 동일한 구간의 원음 신호

※ 본 연구는 방위사업청 및 국방과학연구소의 재원에 의해 설립된 신호정보 특화연구센터 사업의 지원을 받아 수행되었음.
◆ First Author : (ORCID:0000-0001-8626-8355)Department of Electronic Engineering, Seoul National University of Science and Technology, deokkyuyun@gmail.com, 학생회원
° Corresponding Author : (ORCID:0000-0002-5463-1479)Department of Electronic and IT Media Engineering, Seoul National University of Science and Technology, shchoi@seoultech.ac.kr, 정회원
논문번호 : 201903-017-A-LU, Received March 13, 2019; Revised March 21, 2019; Accepted March 25, 2019

를 구성한다. 제안 방법을 통해 충분한 양의 훈련데이터를 구축할 수 있으며, 이는 실제 환경에서의 잡음제거 성능을 높일 수 있다.

Ⅱ. IRM 기반 가상 데이터 생성 모델 및 잡음제거 모델

그림 1과 같이 실제 잡음환경에서 취득한 혼합신호 *y(n)*과 해당 음향의 원음 *x(n)*을 단구간 푸리에 변환(short-time Fourier transform, STFT)하여 *Y(i,k)*와 *X(i,k)*를 구한 뒤 식 (1)과 같은 두 스펙트럼의 비율 *r(i,k)*를 프레임 단위로 훈련시켜 원음으로부터 혼합신호를 생성하는 모델을 구성한다.

$$r(i,k)=\frac{|Y(i,k)|}{|X(i,k)|} \tag{1}$$

여기에서 i,k는 각각 프레임 인덱스와 frequency bin index이며, 가상의 혼합신호 스펙트럼 $|\hat{Y}(i,k)|$은 식 (2)를 통해 생성된다. 여기서 $\hat{r}(i,k)$은 모델의 출력이다.

$$|\hat{Y}(i,k)|=\hat{r}(i,k)|X(i,k)| \tag{2}$$

잡음 제거 모델을 위해서는 그림 1의 가상 데이터 생성모델과 노드 개수, 은닉층 개수, 활성함수 등이 모두 동일한 구조를 갖는 심층신경망에 $|\hat{Y}(i,k)|$를 입력으로, $|X(i,k)|/|\hat{Y}(i,k)|$를 출력으로 하여 잡음제거 모델을 훈련한다.

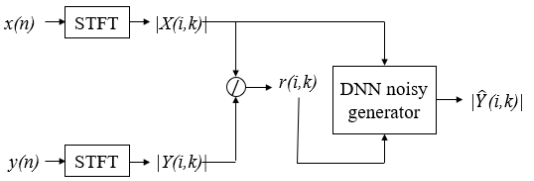

그림 1. 가상 데이터 생성 모델
Fig. 1. Block diagram of virtual noisy generator

Ⅲ. 실험 및 결과

본 논문에서는 STFT를 수행할 때 프레임 길이는 20ms로 하였고, 50% 중첩을 하였다. 샘플링 주파수는 16 kHz로 변환하여 사용하였기 때문에 한 프레임의 샘플 수는 320개이며, 512 포인트로 FFT를 하였고, 257차의 스펙트럼벡터를 사용하였다. DNN모델의 은닉층은 3개이며 각 노드의 수는 1000개이다.

실험을 위해 실제 카페에서 20분 분량의 음악신호 데이터를 취득하였다. 프레임 단위로 계산하면 12만개의 프레임이며, 이중 1/8 분량은 잡음제거 성능 측정을 위한 테스트 데이터(s1~s5)로 사용하였다. 표 1에 '원음에 웅성이는 소리, 커피머신 소리 등을 인위적으로 더하여 구한 데이터로 훈련시킨 잡음제거 모델(Artificial)', '실제 데이터로 훈련시킨 잡음제거 모델(Real)', '제안 방법으로 생성한 데이터로 훈련시킨 잡음제거 모델(Proposed)'의 성능을 식 (3)의 로그 스펙트럼 거리 LSD(Log Spectral Distance)[4]를 사용하여 비교하였으며, 식 (3)에서 $|X(i,k)|$와 $|Y(i,k)|$는 각각 취득한 원음 및 혼합신호의 스펙트럼 크기이다.

$$LSD=\frac{1}{M}\sum_{i=1}^{M}\sqrt{\left(\frac{1}{K}\sum_{k=1}^{K}\left(10\log\left(\frac{|Y(i,k)|^2}{|X(i,k)|^2}\right)\right)^2\right)} \tag{3}$$

표 1에서 알 수 있듯이 잡음을 제거할 대상이 실제 환경에서의 취득한 음성일 경우, 인위적인 덧셈으로 만들어진 기존의 방식으로 훈련된 모델은 성능이 낮다. 제안방법의 경우 실제 환경에서 취득한 신호를 사용하지 않고 가상의 혼합신호를 생성하는 방법임에도 실제 데이터로 훈련한 방식과 비슷한 성능을 나타내며, 기존방법보다 성능이 우수하다.

표 1. 원음과 처리된 신호간의 LSD (dB) 결과
Table 1. LSD (dB) between original signal and processed signal

Method / Signal	Not processed	Artificial	Real	Proposed
s1	12.97	12.02	10.51	10.73
s2	12.57	12.50	11.05	11.35
s3	13.86	11.96	11.20	11.83
s4	15.27	14.05	9.03	10.56
s5	12.39	12.60	9.62	9.95
mean	13.41	12.63	10.28	10.88

Ⅳ. 결론 및 향후 연구방향

본 논문에서는 원음으로부터 실제 잡음환경에서 취득한 혼합신호와 유사한 데이터를 생성하는 딥러닝 기반 가상 데이터베이스 생성 기법을 제안하였다. 이 기법을 통해 시간과 비용을 절약하면서 원음만으로

잡음제거 모델 훈련을 위한 충분한 양의 혼합신호 데이터를 구축할 수 있었고, 잡음제거 모델의 성능도 높일 수 있었다. 향후 제안 기법을 통해 더욱 다양한 원음으로 잡음환경에서의 가상 혼합신호 데이터베이스를 구축하여 잡음제거 성능을 향상시킬 계획이다.

References

[1] Y. Zhao, et al., "DNN-based enhancement of noisy and reverberant speech," *IEEE Int. Conf. Acoustics, Speech and Signal Process,* pp. 6525-6529, Shanghai, China, Mar. 2016.

[2] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," *IEEE Int. Conf. Acoustics, Speech and Signal Process.,* pp. 7092-7096, Vancouver, BC, Canada, May 2013.

[3] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language,* vol. 46, pp. 535-557, Nov. 2017.

[4] A. Gray and J. Markel, "Distance measures for speech processing," *IEEE Trans. Acoustics, Speech, and Signal Process.*, vol. 24, no. 5, pp. 380-391, Oct. 1976.