강화학습을 적용한 EVRC 기반의 음성향상기법에 대한 연구

김소현*, 장준혁**
*한양대학교 전자컴퓨터통신공학과
**한양대학교 융합전자공학부
e-mail : jchang@hanyang.ac.kr

A Study on EVRC-based Speech Enhancement by Reinforcement Learning

Sohyeon Kim*, Joon-Hyuk Chang**
*Dept of Electronics and Computer Engineering, Hanyang University
**School of Electronic Engineering, Hanyang University

요 약

본 논문에서는 음성인식의 성능을 높이기 위해 잡음을 제거하여 음성을 향상시킬 목적으로 심화신경망 기반의 강화학습을 적용한 음성향상 기법을 제안한다. EVRC를 통해 잡음을 제거한 후 강화학습을 적용하여 성능을 비교하며 기존의 음성향상 기법보다 향상된 성능을 가지는 모델을 구현하고자 한다.

1. 서론

효과적인 음성인식을 위해서는 음성신호의 잡음을 제거하는 음성향상 기법이 반드시 필요하다. 과거에는 주로 음성신호와 잡음신호 사의의 통계적 정보를 이용하여 잡음을 제거하는 음성향상기법을 사용하였으나 최근 머신러닝 기법 중 하나인 심화신경망을 적용한 음성향상 기법이 등장하면서 성능의 엄청난 발전을 이루었다. 이에 따라 또 다른 머신러닝 기법인 강화학습을 적용한 음성향상 기법 역시 성능의 향상을 가져올 것으로 기대한다. 본 논문에서는 EVRC를 통한 음성향상 기법에 심화신경망 기반의 강화학습을 적용하여 성능의 향상을 확인한다.

2. 본론

본 논문에서는 EVRC를 통한 음성향상 기법에 심화신경망 기반의 강화학습을 적용하여 음질을 향상시킨 후 PESQ와 인식률을 측정하여 인식 성능을 비교해보았다.

강화학습의 구조는 3개의 은닉층으로 구성되었으며 활성함수는 ReLU를 사용하였다. 학습의 초기단계에서는 EVRC를 통해 잡음을 제거한 학습 데이터셋을 생성한다. 이 때 얻은 음성신호로부터 시간-주파수 마스크를 추출한 후 K-means 군집화를 통해 32개의 시간-주파수 마스크 템플릿들을 생성한다. 또한 MMSE방식을 기반으로 action-value 함수를 선행학습 시킨다. 학습단계의 핵심은 reward를 통해 action-value 함수를 학습시키고 이를 통해 최적의 selection policy를 얻는 것이다. Reward는 해당 policy를 통해 선택한 마스크를 이용해 음성향상 시킨 음성신호의 PESQ를 사용한다. 그런데 PESQ는 SNR, 잡음과 같은 외부적 요인에 영향을 받기 때문에 절대적인 수치를 사용하기 어렵다. 따라서 EVRC를 통해 음성향상 시킨 음성신호의 PESQ 결과와 강화학습을 적용한 EVRC를 통해 음성향상 시킨 음성신호의 PESQ 결과사이의 차를 reward로 정의한다.

$$R = \tanh(\alpha(Z - Z^{EVRC})) \tag{1}$$

여기서 Z는 강화학습을 적용한 EVRC를 통해 얻은 PESQ 값, Z^{EVRC}는 EVRC를 통해 얻은 PESQ값을 의미한다. 그리고 이 때, long-term에서 정의되는 PESQ 수치를 short-term에서 reward로 사용할 수 있도록 (2)-(4)와 같은 추가적인 처리를 거친다.

$$\widetilde{E}_k = \sum_{w=1}^{\Omega} |\ln|Y_{w,k}| - \ln|H_w S_{w,k}||^2 \tag{2}$$

$$E_k = \frac{\widetilde{E}_k}{\max_{k \in K}(\widetilde{E}_k)} \tag{3}$$

$$r_k = \begin{cases} (1-E_k)R & (R>0) \\ E_k R & (\text{그 외}) \end{cases} \tag{4}$$

여기서 $Y_{w,k}$는 음성향상 시킨 음성신호, $H_w S_{w,k}$는 잡음이 없는 깨끗한 음성신호를 의미한다.

강화학습을 적용하지 않은 모델과 비교하여 강화학습을

적용한 EVRC를 이용해 음성 향상시킨 모델의 PESQ 성능이 더 좋을 경우에 reward를 주는 방식, 즉 $R>0$인 경우에 reward를 주는 방식으로 최적의 selection policy를 학습한다.

$$\widetilde{Q}(x_k,a_k)=\begin{cases} r_k+\max_{za\in A}Q(x_k,a) & (R>0) \\ Q(x_k,a_k) & (\text{그 외}) \end{cases} \quad (5)$$

이러한 과정을 반복하며 계속해서 최적의 selection policy를 업데이트하고 강화학습의 출력이 최적의 selection policy와의 차이가 최소가 되도록 학습한다.

학습에 사용된 데이터셋은 총 192000개의 음성신호로 한국어로 녹음된 2000개의 문장에 16가지 잡음환경(babble, white, factory1, hfchannel, music, vacuum, water, animal, cough, door, footstep, refrigerator, shower, snore, TV, wind), 6가지 SNR(0dB, 5dB, 10dB, 15dB, 20dB, 25dB)을 적용하여 사용하였다. 인식 성능의 측정을 위해 한국어로 녹음된 200개의 문장을 테스트 데이터셋으로 사용하였다.

표1. 음성 향상시킨 PESQ 결과

	Babble		White	
	10dB	20dB	10dB	20dB
Noisy	2.32	3.06	1.78	2.89
EVRC	2.56	**3.29**	**2.72**	3.39
proposed	**2.58**	**3.29**	**2.72**	**3.43**

표2. 음성 향상시킨 인식률(WER) 결과

	Babble		White	
	10dB	20dB	10dB	20dB
Noisy	96	62	126	109
EVRC	86	39	111	29
proposed	**80**	**34**	**88**	**28**

실험 결과, PESQ측면에서는 babble 잡음 환경과 white 잡음 환경, 그리고 여러 SNR 상황 모두에서 비슷한 결과를 보였다. 하지만 EVRC에 강화학습을 적용한 모델의 성능이 EVRC를 통해 음성 향상시킨 모델의 성능보다 미세하게 향상된 것을 확인할 수 있었다. WER(Word Error Rate)을 기반으로 한 결과를 보면 PESQ 결과보다 더 확연한 차이를 볼 수 있었다. Babble 잡음 환경과 White 잡음 환경, 그리고 여러 SNR에 대한 모든 상황에 대해서 EVRC에 강화학습을 적용한 모델의 성능이 EVRC를 통해 음성 향상시킨 모델의 성능보다 뛰어났다.

3. 결론

EVRC기반의 음성향상 기법과 EVRC 기반의 음성향상 기법에 심화신경망 기반의 강화학습을 적용한 모델의 성능을 비교한 결과, 심화신경망 기반의 강화학습을 적용한 음성향상 기법이 더 뛰어난 성능을 보였다. 추후 군집화 방법을 다양하게 바꿔 적용해보고 reward를 PESQ가 아닌 다른 수치를 이용하여 계산함으로써 성능을 높일 수 있을 것으로 보인다. 또한 EVRC를 제외한 다른 음성향상 기법에도 심화신경망 기반의 강화학습을 적용하면 향상된 성능을 얻을 수 있을 것으로 보인다.

감사의 글

이 논문은 2017년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행한 연구임 (No.2016-0-00564, 사용자의 의도와 맥락을 이해하는 지능형 인터랙션 기술 연구개발)

참고문헌

[1] Yong Xu, Jun Du, Li-Ring Dai, Chin-Hui Lee, “A regression approach to speech enhancement based on deep neural networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol23, no.1, pp.7-19,2015

[2] Yuma Koizumi, Kenta Niwa, Yusuke Hoika, Kazunori Kobayashi, Yoichi Haneda, “DNN-based source enhancement self-optimized by reinforcement learning using sound quality measurements,” in *Proc. ICASSP*, 2017.