심화 신경망 기반의 음성 향상 기법에 관한 연구

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

A Study on Deep Neural Network based Speech Enhancement

Moa Lee*, Joon-Hyuk Chang**
Dept. of Electronic Engineering, Hanyang University

요 약

본 논문에서는, 적층형 심화 신경망 회귀 모델을 도입하여 잡음이 포함된 입력 신호의 특징벡터로부터 깨끗한 입력 신호의 특징벡터를 추정함으로써 음성 향상 성능을 개선 시켰다. 제안된 방법은 기존의 단일 심화신경망 기법 보다 음성인식 성능 향상에 더욱 효과가 있었다.

1. 서론

사람의 음성과 함께 마이크로 입력되는 배경 잡음은 음성 통신 및 음성인식 등에서 통화 품질이나 음성인식 성능을 저하시키는 주요 원인이 된다. 최근에는 심화 신경망 모델을 음성 향상 기법에 적용하여 성능 저하의 주요 원인이 되는 배경 잡음을 제거하기 위한 연구가 활발히 이루어지고 있다. 음성 향상을 위해 잡음 음성 신호의 특징벡터로부터 깨끗한 음성의 특징벡터를 추정하는 비선형적 회귀모델을 학습하는 기법들이 제안되었다[1][2]. 심화 신경망은 배경 잡음을 제거하기 위한 비선형적 모델을 학습하는데 적용되어 우수한 성능을 보이며 심화 신경망을 통한 잡음 제거의 가능성을 보여주었다.

본 논문에서는 두 개의 비선형적 회귀 모델을 학습하여 순차적으로 쌓은 적층형 심화신경망 기반의 배경 잡음 제거 기법을 제안한다. 심화 신경망 회귀 기법을 적용하여 잡음이 포함된 특징벡터로부터 깨끗한 음성의 특징벡터를 추정하며, 제안하는 기법은 단일 심화 신경망 보다 음성인식 성능에서 우수한 성능을 보였으며, 복잡한 패턴의 잡음 신호로부터 깨끗한 음성 신호를 추정하는데 효과적임을 보였다.

2. 단일 심화신경망 기반의 음성향상 기법 리뷰

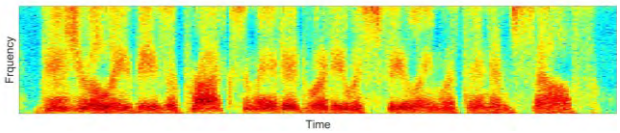
심화신경망 기반의 음성향상 기법에서는 잡음 음성으로부터 추출된 특징벡터로부터 깨끗한 음성의 특징벡터를 타겟으로 추정한다. 이때, 심화신경망은 (1)과 같이 타겟인 깨끗한 음성의 특징벡터와 출력간의 오차를 최소화 하도록 학습된다.

$$E_r = \frac{1}{N}\sum_{n=1}^{N}\left\|\hat{X}(Y_{n-\tau}^{n+\tau}, W, b) - X_n\right\|_2^2 \qquad (1)$$

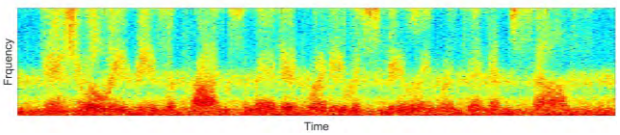
이때 $\hat{X}(Y_{n-\tau}^{n+\tau}, W, b)$과 X_n은 각각 출력과 타겟의 특징벡터이며, N 은 minibatch size 를 나타낸다. Y_n는 입력되는 잡음이 포함된 신호이고, W 와 b 는 학습된 weight 와 bias 파라미터를 나타낸다. 본 논문에서 우리는 음성 특징 벡터로 Mel-frequency cepstral coefficients (MFCC)를 사용하였다.

3. 제안된 적층형 심화신경망 기반의 음성향상 기법

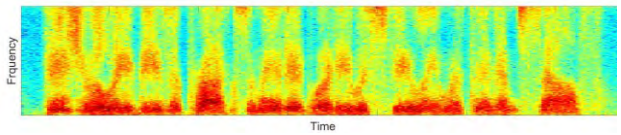
본 논문에서는 음성 향을 위하여 먼저 학습된 단일 심화신경망의 출력을 다시 잡음음성의 특징벡터와 이어 붙여 입력으로 사용하고 깨끗한 음성을 추정하는 심화신경망 모델을 한번 더 학습하여 음성 향상에 사용하였다. 따라서 첫번째 심화신경망 모델에서는 잡음 신호의 특징벡터를 입력으로 받아 깨끗한 음성의 특징 벡터를 추정하고, 두번째 심화신경망 모델에서는 첫번째 심화신경망으로부터 추정된 특징벡터와 잡음 신호의 특징벡터를 다시 이어 붙여 입력으로 하고 깨끗한 음성의 특징벡터를 추정한다. 두개의 심화신경망을 통하여 깨끗한 음성을 추정하므로 잡음 신호의 복잡한 패턴을 보다 더 잘 추정할 수 있다.



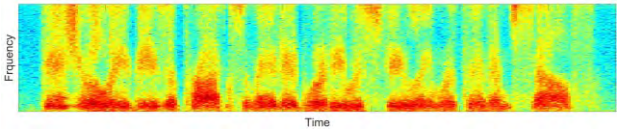
(a) 깨끗한 음성의 스펙트럼



(b) 잡음 신호의 스펙트럼



(c) 단일 심화신경망 기반의 음성향상 기법을 적용한 스펙트럼



(d) 적층형 심화신경망 기반의 음성향상 기법을 적용한 스펙트럼

(그림 1) 스펙트럼 비교

4. 실험 결과 및 비교

제안된 심화 신경망 기반의 음상향상 기법을 평가하기 위하여 TIMIT 데이터 베이스를 이용하였으며, 학습과 테스트시에는 자체적으로 녹음한 배경 잡음이 사용되었다. 심화 신경망은 모두 3 개의 은닉층과 1024 개의 은닉노드로 구성되었다. 제안된 기법은 기존의 단일 심화 신경망 기반의 기법과 비교되었으며, 표 1 은 PER (Phone error rate) 결과를 보여준다.

<표 1> 단일 심화신경망과 적층형 심화신경망 기반의 음상향상 기법을 적용한 음성인식에서의 PER(Phone error rate) 비교

음성 향상 기법 모델	PER (%)				
	5dB	10dB	15dB	20dB	Avg.
사용 안함	27.7	25.2	24.4	23.8	25.28
단일 모델	27.0	24.9	23.3	22.7	24.48
적층형 모델	26.4	24.6	23.1	22.5	24.15

5. 결론

본 논문에서는 심화 신경망 기반의 배경잡음 제거 기법을 제안하였다. 제안된 적층형 심화신경망 기반은 잡음 제거 기법은 기존의 단일 심화신경망 기반의 잡음 제거 기법보다 PER 면에서 우수한 성능을 보였다.

ACKNOWLEDGMENT

이 연구는 방위사업청 및 국방과학연구소의 재원에 의해 설립된 신호정보 특화연구센터 사업의 지원을 받아 수행되었음.

참고문헌

[1] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “An experimental study on speech enhancement based on deep neural networks,”IEEE Signal Process. Lett., vol. 21, no. 1, pp. 65-68, Jan. 2014.

[2] C. M. Lee, J. W. Shin, and N. S. Kim, "DNN-based residual echo suppression," in Proc. Interspeech, Sep. 2015, pp. 1775-1779.