Journal of Korea Multimedia Society Vol. 22, No. 7, July 2019(pp. 749-758)
https://doi.org/10.9717/kmms.2019.22.7.749

딥 뉴럴 네트워크 기반의 음성 향상을 위한 데이터 증강

이승관†, 이상민††

Data Augmentation for DNN-based Speech Enhancement

Seung Gwan Lee†, Sangmin Lee††

ABSTRACT

This paper proposes a data augmentation algorithm to improve the performance of DNN(Deep Neural Network) based speech enhancement. Many deep learning models are exploring algorithms to maximize the performance in limited amount of data. The most commonly used algorithm is the data augmentation which is the technique artificially increases the amount of data. For the effective data augmentation algorithm, we used a formant enhancement method that assign the different weights to the formant frequencies. The DNN model which is trained using the proposed data augmentation algorithm was evaluated in various noise environments. The speech enhancement performance of the DNN model with the proposed data augmentation algorithm was compared with the algorithms which are the DNN model with the conventional data augmentation and without the data augmentation. As a result, the proposed data augmentation algorithm showed the higher speech enhancement performance than the other algorithms.

Key words: Speech Enhancement, Data Augmentation, Deep Neural Network(DNN), Noise Reduction

1. 서 론

휴대폰, 전화기, 음성 인식 스피커, 보청기 등 음성과 관련된 장치들은 주변 잡음에 따라 성능이 저하될 수 있기 때문에 잡음을 제거하는 전처리 기술인 음성 향상 기술의 사용이 필수적이다. 때문에 스펙트럼 차감법(spectral subtraction)[1]부터 위너 필터링(Wiener filtering)[2], 서브스페이스 방법(subspace methods)[3], 칼만 필터링(Kalman filtering)[4] 까지 다양한 음성향상 알고리즘들이 연구되어 왔다. 최근에는 딥 뉴럴 네트워크(deep neural network)를 기반으로 하는 음성 향상 알고리즘[5]이 우수한 성과를 보여주고 있다.

잘 훈련된 딥 뉴럴 네트워크 모델을 만들기 위해서는 많은 양의 훈련 데이터를 필요로 하지만 물리적으로 데이터의 양은 한정되어 있으며, 데이터의 수집에는 많은 비용과 시간을 필요로 한다. 때문에 데이터가 한정된 상황에서 성능을 높일 수 있는 다양한 기술들이 연구되고 있으며, 가장 대표적으로 데이터를 가공하여 데이터의 양을 증대시키는 방법인 데이터 증강(data augmentation) 방법이 있다. 본 논문에서는 음성의 공명 주파수이자 언어의 명료도를 관장

※ Corresponding Author : Sangmin Lee, Address: (22212) Inha Univesity, 100, Inha-ro, Nam-gu, Incheon, Republic of Korea, TEL : +82-32-860-7420, FAX : +82-32-860-1333, E-mail : sanglee@inha.ac.kr
Receipt date : Apr. 5, 2019, Revision date : May 28, 2019
Approval date : June 10, 2019

† Dept. of Electronic Engineering, Inha University (E-mail : lsg2578@hanmail.net)
†† Dept. of Electronic Engineering, Inha University
※ This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (2016R1A2B4015370)

하는 포먼트 주파수에 가중치를 부여하는 포먼트 강화(formant enhancement)를 이용하여 훈련 데이터의 양을 증대시키는 데이터 증강 알고리즘을 연구하였다.

제안한 데이터 증강 알고리즘을 사용하여 훈련된 딥 뉴럴 네트워크 모델은 객관적 음질 평가 방법인 ITU-T P.862 PESQ(Perceptual Evaluation of Speech Quality)[6]와 명료도 평가 방법인 STOI (Short Time Objective Intelligibility)[7]를 사용하여 음성 향상 성능을 평가 하였다. 그 결과 본 논문에서 제안한 포먼트 강화를 이용한 데이터 증강 알고리즘을 사용한 경우 기존의 데이터 증강 알고리즘과 데이터 증강 알고리즘을 사용하지 않은 경우와 비교해 더 높은 음성 향상 성능을 보여주었다.

2. 이 론

2.1 딥 뉴럴 네트워크 기반의 음성 향상

딥 뉴럴 네트워크 기반의 음성 향상 방법은 Fig. 1에서 확인 할 수 있듯이 크게 훈련을 하는 훈련 단계(training stage)와 훈련된 모델을 기반으로 노이즈를 제거하여 음성 품질을 향상 시키는 향상 단계(enhancement stage)로 나누어진다[5].

훈련 단계에서는 잡음이 존재하는 음성 신호와 잡음이 존재하지 않는 깨끗한 음성 신호를 각각 입력 특징 벡터와 목표 특징 벡터로 사용하여 매핑 기능을 훈련하며, 평균제곱오차(mean square error) 손실 함수를 사용하여 보다 정확하게 잡음을 추정 할 수 있도록 훈련시킨다. 향상 단계에서는 앞서 훈련시킨 딥 뉴럴 네트워크 모델을 사용하여 잡음이 존재하는 음성에서 잡음을 추정하고 제거하여 향상된 음성 신호를 얻을 수 있다.

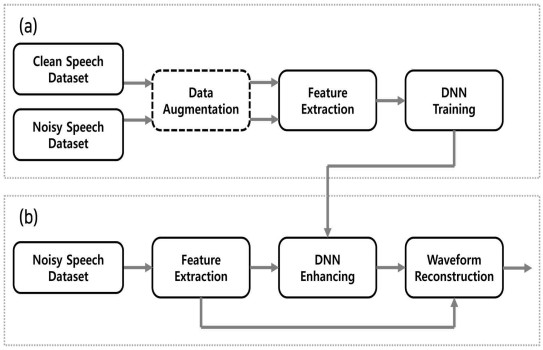

Fig. 1. A block diagram of the speech enhancement based on DNN. (a) Training stage, (B) Enhancement stage.

본 논문에서는 딥 뉴럴 네트워크 모델의 훈련 단계에서 포먼트 강화를 이용한 데이터 증강 알고리즘을 사용하여 음성 향상의 성능을 향상 시킬 수 있는 방법을 제안하였다.

2.2 기존의 데이터 증강 알고리즘

딥 뉴럴 네트워크의 성능을 높이기 위해서는 충분한 양의 훈련 데이터가 필수적이다. 하지만 물리적으로 데이터의 양은 한정되어 있으며 더 많은 데이터의 수집은 곧 많은 시간과 비용이 소요됨을 의미한다. 때문에 이미지 분류부터 음성과 음악 분류, 의학 이미지 분류[8] 까지 다양한 딥 러닝 분야에서 훈련 데이터를 가공하여 유사한 데이터를 생성하는 데이터 증강 방법을 사용하여 딥 뉴럴 네트워크의 성능을 높이고 있다. Fig. 2는 이미지 분류를 위한 딥 뉴럴 네트워크에서 주로 사용하는 데이터 증강 알고리즘의 예시를 나타내고 있다. 1개의 자동차 이미지 데이터에 다양한 효과를 적용하여 원본 대비 5배로 증강 시켰다.

음성인식 또는 음성분류 같은 음성 딥러닝 분야에서의 데이터 증강 방법으로는 VTLP(Vocal Track Length Perturbation)[9], 피치 변경(pitch perturbation)과 다이나믹레인지 변경(dynamic range perturbation)[10], 속도 변경(speed perturbation)과 템포 변경(tempo perturbation)[11] 등의 다양한 방법들이 사용되어 왔다. 다수 논문의 실험 결과에 따르면 다음과 같은 두 가지 알고리즘이 다른 알고리즘에 비해 높은 성능을 보여주었다.

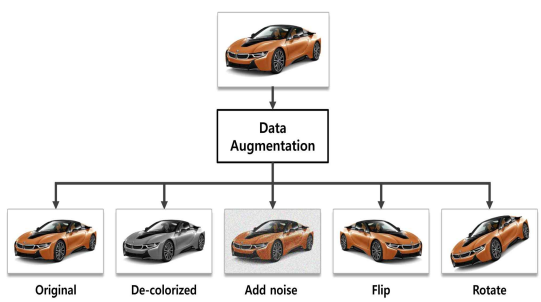

Fig. 2. Example of data augmentation method using car image.

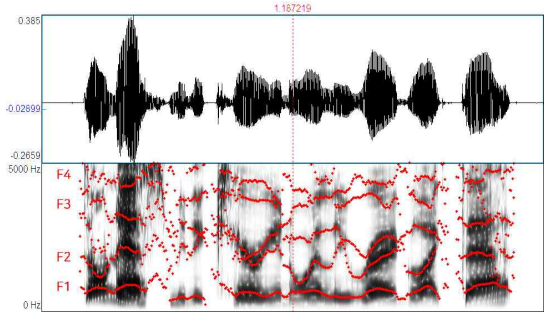

Fig. 3. Formant on speech waveform and spectrogram.

2.2.1 속도 변경 기반의 데이터 증강

음성 신호를 재 샘플링 하여 신호의 속도를 수정하는 방법이다. 사운드 프로세싱 프로그램인 SoX 툴을 사용하여 구현이 가능하다. 속도를 원래 속도의 90%와 110%로 수정한 데이터를 추가하여 원본 대비 3배로 증강 시킨 경우 가장 좋은 성능을 보여주었다.

2.2.2 템포 변경 기반의 데이터 증강

음성 신호의 피치와 스펙트럼이 변하지 않도록 주의하면서 신호의 템포 즉 말의 속도를 수정하여 데이터를 증강 시키는 방법이다. 속도 변화 방법과 동일하게 SoX 툴을 사용하여 구현이 가능하다. 템포를 원래 속도의 90%와 110%로 수정한 데이터를 추가하여 원본 대비 3배로 증강 시킨 경우 가장 띄어난 성능을 보여주었다.

2.3 포먼트 강화

인간의 성대에서 형성된 음성은 성도(vocal tract)를 통과하면서 공명이나 간섭 등의 영향을 받게 되어 진폭이 일정하게 커지고 작아지는 것을 반복한다. 이러한 일련의 변화에서 진폭과 에너지가 높은 정점을 포먼트 주파수(formant frequency)[12]라 하며, 만들어진 순서에 따라 F1, F2, F3, ... , Fn으로 표현한다. 사람에 따라서 조금씩 차이가 있지만 일반적으로 3-5개 정도의 포먼트가 형성된다. Fig. 3은 특정 화자의 발화 파형과 스펙트로그램을 나타내고 있다. 특정 주파수 대역에 형성되어 있는 에너지군인 포먼트를 확인 할 수 있다.

Table 1은 Peterson & Barney가 총 76명의 사람을 대상으로 연구한 영어 모음과 성별에 따른 포먼트 주파수를 나타내고 있다[13]. F1 주파수는 270-860 Hz 구간에, F2 주파수는 840-2790Hz 구간에, F3 주파수는 2240-3310Hz 구간에 분포되어 있는 것을 확인 할 수 있다. F1, F2 포먼트는 모음을 인식하는데 영향을 미치기 때문에 모음 음형대(vowel formant)라고 부르며, F3 이후의 포먼트들은 목소리의 특성에 영향을 미치기 때문에 가수 음형대(singer's formant)라고 부른다.

보청기에서는 음성의 인지도를 높이기 위하여 음성 데이터에서 포먼트를 찾아내어 가중치를 부가하는 포먼트 강화(formant enhancement)를 사용한다. 본 논문에서는 Peterson & Barney가 연구한 F1, F2 포먼트 주파수 구간을 기반으로 포먼트 강화를 사용하여 데이터를 증강하는 방법을 연구하였다.

Table 1. F1, F2, F3 frequencies by gender and English vowels studied by Peterson & Barney

	F1		F2		F3	
	M	W	M	W	M	W
i	**270**	310	2290	**2790**	3010	**3310**
I	390	430	1990	2480	2550	3070
3	530	610	1840	2330	2480	2990
æ	660	**860**	1720	2050	2410	2850
a	730	850	1090	1220	2440	2810
ɔ	570	590	**840**	920	2410	2710
U	440	470	1020	1160	**2240**	2680
u	300	370	870	950	2240	2670
ʌ	640	760	1190	1400	2390	2780
3˜	490	500	1350	1640	1690	1960

3. 제안한 방법

3.1 제안한 포먼트 강화를 이용한 데이터 증강 알고리즘

첫 번째로 데이터 증강에 사용할 포먼트를 선택하였다. F3는 음소의 음질과 음색을 결정하지만 F1, F2에 비해 언어의 인지에 미치는 영향은 크지 않다고 알려져 왔다. F3의 사용 여부를 결정하기 위하여 F1, F2를 증강시켜 원본 대비 3배로 증강시킨 경우와, F1, F2, F3를 증강시켜 원본 대비 4배로 증강시킨 경우로 나누어 딥 뉴럴 네트워크를 각각 훈련시킨 후 성능을 비교해 보았다. 그 결과 Table 2에서 확인할 수 있듯이 F1, F2를 증강시킨 경우가 F1, F2, F3를 증강시킨 경우보다 총 6개 잡음 중 5개 잡음에서 더 높은 잡음 제거 성능을 보였다.

두 번째로 F1, F2 포먼트 강화를 위하여 적용할 최적의 가중치를 연구하였다. 데이터 증강 알고리즘에서 너무 미세한 변화는 효과가 없고, 큰 변화는 도리어 왜곡으로 작용해 오히려 성능을 떨어뜨리는 원인이 될 수 있다. 때문에 가장 효과적으로 성능을 높일 수 있는 가중치를 찾는 것이 중요하다. 가장 효과적인 가중치를 찾기 위하여 F1, F2 포먼트 주파수 구간에 1.3에서 3.5까지 다양한 가중치를 적용한 훈련 데이터를 준비하여 딥 뉴럴 네트워크를 훈련 후 성능을 비교하였다. 그 결과 Table 3에서 확인 할 수 있듯이 1.5의 가중치를 적용한 경우 다른 가중치에 비해 높은 향상율을 확인 할 수 있었다.

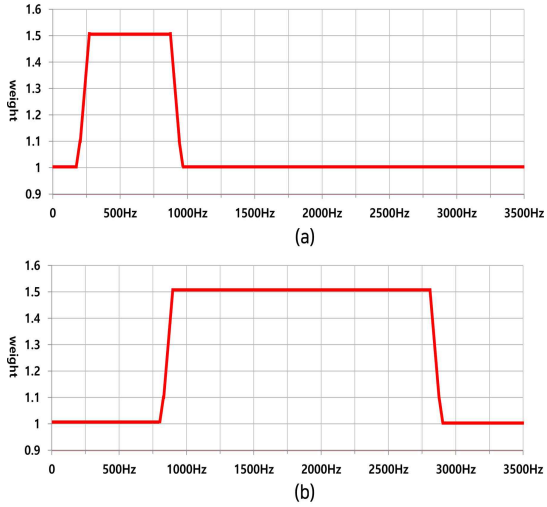

Fig. 4. The weight applied to the formant frequency. (a) Weight for F1 formant enhancement, (b) Weight for F2 formant enhancement

Fig. 4는 포먼트 주파수에 최종적으로 적용된 가중치를 나타내고 있다. 포먼트 향상을 위해 필터 뱅크를 사용하여 F1에 해당하는 270-860Hz 구간에 1.5의 가중치를 부여 하였으며, F2에 해당하는 840-2790Hz 구간에도 1.5의 가중치를 적용 하였다. 갑작스러운 변화로 인한 데이터의 왜곡이 발생하는 것을 방지하기 위하여 경계 부분의 주파수 대역에도 1.0에서 1.5의 가중치를 적용 하였다.

Table 2. The average PESQ improvement rate of the enhanced speech by using F1, F2 augmentation and F1, F2, F3 augmentation.

	babble	buccaneer	factory	leopard	volvo	white
F1, F2	**22.64%**	**36.73%**	**26.98%**	**28.49%**	**10.04%**	44.47%
F1, F2, F3	18.95%	34.26%	29.16%	15.64%	9.67%	**49.19%**

Table 3. The average PESQ improvement rate by weight

	weight 1.3	weight 1.5	weight 1.7	weight 2.0	weight 2.5	weight 3.5
babble	19.96%	**22.64%**	20.23%	17.24%	19.90%	19.91%
buccaneer	35.01%	**36.73%**	35.38%	34.56%	30.70%	26.80%
factory	24.03%	**27.11%**	26.98%	26.88%	19.87%	18.70%
leopard	26.87%	**28.49%**	28.13%	27.84%	24.30%	22.47%
volvo	9.72%	**10.04%**	10.02%	9.54%	7.42%	7.20%
white	39.46%	44.47%	44.82%	**45.04%**	25.61%	20.61%
avg.	25.84%	**28.73%**	27.59%	26.85%	21.30%	19.28%

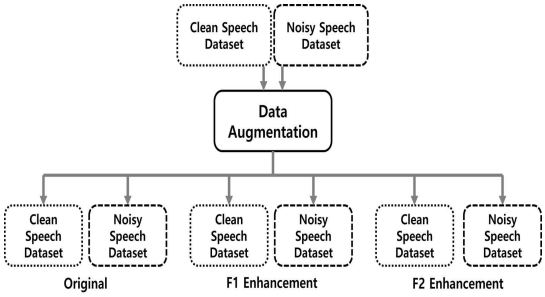

Fig. 5. A block diagram of data augmentation using the formant enhancement.

Fig. 5는 최종적으로 제안한 포먼트 강화를 이용한 데이터 증강 알고리즘의 블록도를 나타내고 있다. 딥 뉴럴 네트워크의 훈련을 위하여 준비된 잡음이 존재하지 않는 깨끗한 음성 데이터와 잡음이 존재하는 음성 데이터에 F1가 강화된 데이터와 F2가 강화된 데이터가 각각 더해져 전체 데이터의 양은 각각 원본 대비 3배로 증가하였다.

3.2 실험 방법

본 논문에서 제안한 데이터 증강 알고리즘을 사용한 훈련 데이터, 기존의 음성 딥 러닝 분야에서 사용된 데이터 증강 알고리즘을 사용한 훈련 데이터, 데이터 증강 알고리즘을 사용하지 않은 훈련 데이터를 사용하여 딥 뉴럴 네트워크 모델을 훈련 후 테스트 데이터를 사용하여 음성 향상 성능을 각각 비교하였다.

실험은 다양한 딥 뉴럴 네트워크 모델을 사용하여 훈련과 평가가 가능한 speech enhancement toolkit [14]을 기반으로 진행하였다. 딥 뉴럴 네트워크 모델로는 최근 음성 향상과 음성 인식 등의 분야에서 높은 성과를 보여주고 있는 LSTM(Long Short Term Memory)[15]을 사용하여 훈련하였다. LSTM은 RNN(Recurrent Neural Network)의 변형 모델로서 메모리 셀(memory sell)과 셀 스테이트(sell state)을 사용하여 오래된 정보를 지속적으로 유지할 수 있도록 설계되었다. 때문에 학습 초기에 입력된 정보가 학습의 마지막까지 영향을 줄 수 있는 이점이 있다.

3.2.1 훈련 데이터

딥 뉴럴 네트워크의 훈련을 위해 약 6시간(7,000문장) 분량의 잡음이 존재하지 않는 깨끗한 음성 데이터와 잡음이 존재하는 음성 데이터를 각각 준비하였다. 잡음이 존재하지 않는 깨끗한 음성 데이터는 TIMIT[16] 음성 데이터를 그대로 사용 하였으며, 잡음이 존재하는 음성 데이터의 경우 TIMIT 음성 데이터에 NOISEX-92[17]의 잡음 데이터를 이용하여 음성 신호를 오염시켰다. 먼저 NOISEX-92의 14가지 잡음 데이터에서 무작위로 1개의 잡음을 선택 한 후 -5dB, 0dB, 5dB, 10dB, 15dB, 20dB 중 무작위로 1개의 SNR 레벨을 선택하여 오염시켰다. 이후 본 논문에서 제안한 포먼트 향상을 이용한 데이터 증강 알고리즘과 기존 음성 딥 러닝 분야에서 사용된 속도 변경, 템포 변경을 이용한 데이터 증강 알고리즘을 사용한 훈련 데이터를 생성하였다.

Table 4는 딥 뉴럴 네트워크의 훈련을 위해 최종적으로 준비된 4가지 훈련 데이터를 나타내고 있다. 첫 번째로 데이터 증강을 사용하지 않은 훈련 데이터가 준비되었다. 약 6시간(7,000문장) 분량의 깨끗한 음성 데이터와 잡음이 존재하는 음성 데이터가 각각 준비되었다. 두 번째로 본 논문에서 제안한 포먼트 향상을 이용한 데이터 증강 알고리즘을 사용한 훈련 데이터를 제작하였다. F1와 F2 포먼트 향상을 통하여 원본데이터 대비 3배 증강된 약 18시간(21,000문장) 분량의 깨끗한 음성 데이터와 잡음이 존재하는 음성 데이터가 각각 준비되었다. 세 번째로 속도 변경을 이용한 데이터 증강 알고리즘을 사용한 훈련 데이터를 제작하였다. 속도를 원본 대비 90%와 110%로 변경하여 원본 데이터 대비 3배 증강된 약 18시간(21,000문장) 분량의 데이터가 준비되었다. 마지막으로 템포 변경을 이용한 데이터 증강 알고리즘의 사용한 훈련 데이터를 제작하였다. 템포를 원본 대비 90%와 110%로 변경하여 원본 데이터 대비 3배 증강된 약 18시간(21,000문장) 분량의 데이터가 준비되었다.

3.2.2 테스트 데이터

훈련이 끝난 모델의 잡음 제거 성능 테스트를 위하여 TIMIT 데이터를 NOISEX-92의 잡음 데이터로 오염시켰다. 먼저 TIMIT에서 10개의 문장을 무작위로 선택한 후 babble, factory, volvo, white 등 14개의 잡음을 0dB, 5dB, 10dB의 3가지 SNR 레벨로 적용시켜 총 420개의 테스트 데이터를 준비하였다.

3.2.3 객관적 음질 및 명료도 평가 방법

본 논문에서 제안한 알고리즘과 기존 알고리즘을

Table 4. Four training data sets prepared for the training of DNN

Dataset	Dataset type	Amount	Total Amount
without data augmentation	original data	clean 6h noisy 6h	clean 6h noisy 6h
proposed data augmentation (formant enhancement)	original data	clean 6h noisy 6h	clean 18h noisy 18h
	F1 enhancement	clean 6h noisy 6h	
	F2 enhancement	clean 6h noisy 6h	
existing data augmentation (speed perturbation)	original data	clean 6h noisy 6h	clean 18h noisy 18h
	90% of original speed	clean 6h noisy 6h	
	110% of original speed	clean 6h noisy 6h	
existing data augmentation (tempo perturbation)	original data	clean 6h noisy 6h	clean 18h noisy 18h
	90% of original tempo	clean 6h noisy 6h	
	110% of original tempo	clean 6h noisy 6h	

성능을 평가하기 위하여 가장 대표적인 객관적 음질 평가 방법인 ITU-T P.862 PESQ(Perceptual Evaluation of Speech Quality)[6]와 명료도의 평가 방법인 STOI(Short Time Objective Intelligibility)[7]를 사용하였다.

PESQ는 인간의 지각 요소를 기초로 한 주관적인 음질을 객관적 수치로 평가 할 수 있기 때문에 음질을 평가하는 객관적인 방법으로 널리 사용되고 있다. 점수의 범위는 -0.5에서 4.5까지의 값을 가지게 된다. 점수가 4.5에 가까울수록 음질이 좋음을 의미하며, -0.5에 가까울수록 음질이 나쁨을 의미한다.

STOI는 기준 신호와 테스트 신호의 짧은 구간을 시간-주파수 영역에서 주파수 가중치를 두어 상관도를 계산하는 방법으로 인간의 음성 청취 관점에서 평가되는 음성의 명료도와 높은 상관관계가 있는 것으로 알려져 있다. 점수의 범위는 0에서 1까지의 값을 가지게 된다. 1에 가까울수록 음성의 명료도가 좋음을 의미하고, 0에 가까울수록 음성의 명료도가 나쁨을 의미한다.

4. 실험 결과 및 고찰

Table 5는 잡음으로 오염된 음성신호, 데이터 증강 없이 딥 뉴럴 네트워크를 사용하여 음성 향상된 신호, 본 논문에서 제안한 포먼트 향상을 이용한 데이터 증강 알고리즘을 사용하여 음성 향상된 신호, 속도 변경을 이용한 데이터 증강 알고리즘을 사용하여 음성 향상된 신호, 템포 변경을 이용한 데이터 증강 알고리즘을 사용하여 음성 향상된 신호의 잡음별 PESQ 점수를 나타낸다. 결과를 살펴보면 본 논문에서 제안한 포먼트 향상을 이용한 데이터 증강 알고리즘을 사용한 경우 14개 잡음 중 12개의 잡음에서, 속도 변경과 템포 변경을 이용한 데이터 증강 알고리즘의 경우 각각 1개의 잡음에서 다른 알고리즘에 비해 우수한 잡음 제거 성능을 보여 주었다. 전체 실험에서 제안한 알고리즘의 평균 PESQ 향상율은 27.4%로 가장 높은 성능을 보여 주었으며, 템포 변경을 이용한 알고리즘의 경우 23.4%, 속도 변경을 이용한 알고리즘의 경우 21.7%, 데이터 증강을 사용하지 않은 경우 21.2%로 그 뒤를 이었다.

Table 5. Evaluation results using PESQ

Noise type		Noisy	Without Data Augmentation	Data Augmentation (Speed)	Data Augmentation (Tempro)	Data Augmentation (Formant)
Babble	0dB	1.7154	1.9882	2.0623	2.0419	**2.1431**
	5dB	2.0600	2.3860	2.4897	2.4734	**2.5230**
	10dB	2.4360	2.7804	2.8707	2.8607	**2.9322**
	avg.	2.0705	2.3849	2.4742	2.4587	**2.5328**
Buccaneer1	0dB	1.5506	2.0646	1.9398	2.0601	**2.1659**
	5dB	1.8458	2.4318	2.3885	2.4636	**2.5321**
	10dB	2.2009	2.8227	2.8588	2.8828	**2.8882**
	avg.	1.8658	2.4397	2.3957	2.4688	**2.5287**
Buccaneer2	0dB	1.4453	1.7184	1.5840	1.6684	**1.9581**
	5dB	1.8173	2.2381	2.0743	2.1420	**2.3240**
	10dB	2.1938	2.7059	2.5432	2.6445	**2.7687**
	avg.	1.8188	2.2208	2.0672	2.1517	**2.3503**
Destroyer engine	0dB	1.7320	2.0312	**2.1817**	2.1359	2.1253
	5dB	2.0121	2.4011	**2.5783**	2.5394	2.5004
	10dB	2.3194	2.7797	**2.9572**	2.9204	2.8422
	avg.	2.0212	2.4040	**2.5724**	2.5319	2.4893
Destroyer ops	0dB	1.8140	2.1629	2.2306	2.2573	**2.3442**
	5dB	2.1649	2.5049	2.6552	2.6427	**2.7535**
	10dB	2.5195	2.8469	2.9808	2.9656	**3.0646**
	avg.	2.1662	2.5049	2.6222	2.6219	**2.7208**
F16	0dB	1.7340	2.1703	2.1180	2.1704	**2.2622**
	5dB	2.0604	2.5523	2.5089	2.5745	**2.5874**
	10dB	2.4361	2.9506	2.8839	2.9507	**2.9657**
	avg.	2.0768	2.5577	2.5036	2.5652	**2.6051**
Factory1	0dB	1.7064	2.1479	2.1092	2.1486	**2.1503**
	5dB	2.1024	2.6373	2.6700	2.7110	**2.7212**
	10dB	2.4090	2.9776	3.0414	3.0550	**3.0688**
	avg.	2.0726	2.5876	2.6069	2.6382	**2.6468**
Factory2	0dB	1.9288	2.4731	2.5873	2.5620	**2.6128**
	5dB	2.3588	2.8912	2.9504	2.9612	**3.0090**
	10dB	2.6834	3.1672	3.1994	3.2423	**3.2588**
	avg.	2.3237	2.8438	2.9124	2.9218	**2.9602**
Leopard	0dB	2.2596	2.4822	2.7433	2.6514	**3.0281**
	5dB	2.5799	2.7846	3.0161	2.9059	**3.2939**
	10dB	2.8541	3.0679	3.2922	3.1886	**3.5864**
	avg.	2.5645	2.7782	3.0172	2.9153	**3.3028**
M109	0dB	2.0784	2.4688	2.6719	2.6161	**2.7123**
	5dB	2.4048	2.8162	3.0129	2.9631	**3.0278**
	10dB	2.7514	3.1927	3.3399	3.2928	**3.3474**
	avg.	2.4116	2.8259	3.0083	2.9573	**3.0292**
Machinegun	0dB	2.4074	2.7424	2.9439	2.8980	**3.0277**
	5dB	2.6924	2.9341	3.1133	3.0754	**3.1792**
	10dB	2.9880	3.1569	3.2913	3.2558	**3.4139**
	avg.	2.6959	2.9445	3.1162	3.0764	**3.2069**
Pink	0dB	1.6075	2.1752	1.9655	2.1734	**2.2265**
	5dB	1.9698	2.6304	2.5133	2.6343	**2.6402**
	10dB	2.3298	3.0522	2.9520	3.0606	**3.0656**
	avg.	1.9690	2.6192	2.4770	2.6228	**2.6441**
Volvo	0dB	3.1017	3.6526	3.6480	**3.6600**	3.6510
	5dB	3.4660	3.8284	**3.8499**	3.8451	3.8040
	10dB	3.8017	3.9659	**3.9765**	3.9709	3.8699
	avg.	3.4564	3.8156	3.8248	**3.8254**	3.7750
White	0dB	1.4333	2.1061	1.7654	2.0264	**2.1760**
	5dB	1.7376	2.4822	2.2295	2.4278	**2.5471**
	10dB	2.0812	2.8697	2.6653	2.8319	**2.8802**
	avg.	1.7507	2.4860	2.2201	2.4287	**2.5344**

Table 6. Evaluation results using STOI

Noise type		Noisy	Without Data Augmentation	Data Augmentation (Speed)	Data Augmentation (Tempro)	Data Augmentation (Formant)
Babble	0dB	0.6370	0.6651	0.6791	0.6814	**0.6878**
	5dB	0.7628	0.7968	0.8019	0.8024	**0.8116**
	10dB	0.8544	0.8748	0.8759	0.8757	**0.8831**
	avg.	0.7514	0.7789	0.7856	0.7865	**0.7942**
Buccaneer1	0dB	0.6156	0.7006	0.6843	0.7056	**0.7096**
	5dB	0.7355	0.8020	0.7944	0.8048	**0.8057**
	10dB	0.8436	0.8809	0.8720	0.8810	**0.8891**
	avg.	0.7316	0.7945	0.7836	0.7971	**0.8014**
Buccaneer2	0dB	0.6238	0.6451	0.6350	0.6475	**0.6719**
	5dB	0.7372	0.7688	0.7577	0.7774	**0.7855**
	10dB	0.8432	0.8634	0.8487	0.8703	**0.8734**
	avg.	0.7347	0.7591	0.7472	0.7651	**0.7770**
Destroyer engine	0dB	0.6680	0.7087	0.7287	0.7269	**0.7343**
	5dB	0.7916	0.8359	0.8380	0.8383	**0.8438**
	10dB	0.8851	0.9043	0.9001	0.9015	**0.9061**
	avg.	0.7816	0.8163	0.8223	0.8222	**0.8281**
Destroyer ops	0dB	0.6960	0.7320	0.7481	0.7572	**0.7708**
	5dB	0.7840	0.8140	0.8256	0.8298	**0.8392**
	10dB	0.8614	0.8760	0.8782	0.8822	**0.8847**
	avg.	0.7805	0.8074	0.8173	0.8231	**0.8316**
F16	0dB	0.6859	0.7050	0.7128	0.7235	**0.7333**
	5dB	0.7888	0.8132	0.8143	0.8283	**0.8296**
	10dB	0.8848	0.8989	0.8911	0.8996	**0.9006**
	avg.	0.7865	0.8057	0.8061	0.8171	**0.8212**
Factory1	0dB	0.6692	0.7048	0.6950	**0.7286**	0.6944
	5dB	0.7699	0.8213	0.8171	**0.8319**	0.8275
	10dB	0.8582	0.8827	0.8749	**0.8828**	**0.8828**
	avg.	0.7658	0.8029	0.7957	**0.8145**	0.8016
Factory2	0dB	0.7392	0.7893	0.7922	0.8025	**0.8083**
	5dB	0.8418	0.8713	0.8625	0.8703	**0.8742**
	10dB	0.9104	0.9127	0.9067	0.9104	**0.9138**
	avg.	0.8305	0.8577	0.8538	0.8611	**0.8654**
Leopard	0dB	0.8015	0.8310	0.8456	0.8460	**0.8534**
	5dB	0.8416	0.8663	0.8750	0.8724	**0.8771**
	10dB	0.8760	0.8913	0.8972	0.8935	**0.8998**
	avg.	0.8397	0.8629	0.8726	0.8707	**0.8768**
M109	0dB	0.7662	0.8019	0.8154	0.8143	**0.8235**
	5dB	0.8440	0.8759	0.8754	0.8773	**0.8785**
	10dB	0.9047	0.9136	0.9082	0.9129	**0.9142**
	avg.	0.8383	0.8638	0.8663	0.8682	**0.8721**
Machinegun	0dB	0.8066	0.8438	0.8658	0.8635	**0.8724**
	5dB	0.8603	0.8750	0.8898	0.8861	**0.8949**
	10dB	0.9001	0.9050	0.9085	0.9093	**0.9187**
	avg.	0.8557	0.8746	0.8880	0.8863	**0.8953**
Pink	0dB	0.6573	0.7081	0.6734	**0.7430**	0.7201
	5dB	0.7777	0.8195	0.8033	**0.8323**	0.8140
	10dB	0.8704	0.8848	0.8760	**0.8895**	0.8833
	avg.	0.7685	0.8041	0.7842	**0.8216**	0.8058
Volvo	0dB	0.9172	**0.9216**	0.9099	0.9184	0.9156
	5dB	**0.9448**	0.9353	0.9284	0.9341	0.9315
	10dB	**0.9654**	0.9436	0.9384	0.9438	0.9418
	avg.	**0.9425**	0.9335	0.9256	0.9321	0.9296
White	0dB	0.6517	0.7104	0.6527	**0.7359**	0.7245
	5dB	0.7654	0.7976	0.7724	**0.8179**	0.8069
	10dB	0.8618	0.8691	0.8520	**0.8810**	0.8735
	avg.	0.7596	0.7923	0.7591	**0.8116**	0.8016

Table 6은 각 신호별 STOI 점수를 나타낸다. 결과를 살펴보면 본 논문에서 제안한 포먼트 향상을 이용한 데이터 증강 알고리즘을 사용한 경우 14개 잡음 중 9개의 잡음에서, 템포 변경을 이용한 데이터 증강 알고리즘의 경우 3개의 잡음에서 다른 알고리즘에 비해 우수한 잡음 제거 성능을 보여 주었다. machinegun 잡음의 경우 모든 알고리즘에서 음성 향상 이후 STOI 점수가 오히려 악화되었는데 짧은 구간을 분할하여 평가하는 방법상의 한계로 인한 문제로 추정된다.

전체 실험에서 제안한 알고리즘의 평균 STOI 향상율은 5.2%로 가장 높은 성능을 보여 주었으며, 템포 변경을 이용한 경우 4.9%, 데이터 증강을 사용하지 않은 경우 3.7%, 속도 변경을 이용한 경우 3.3%, 로 그 뒤를 이었다.

5. 결 론

본 논문에서는 음성의 공명 주파수이자 언어의 명료도를 영향을 미치는 포먼트 주파수에 가중치를 부여하는 포먼트 강화를 이용하여 훈련 데이터의 양을 증대시키는 데이터 증강 알고리즘이 제안하였다. 결과적으로 PESQ 테스트에서는 총 14개중 12개의 잡음 환경에서, STOI 테스트에서는 총 14개중 9개의 잡음 환경에서 데이터 증강 알고리즘 사용하지 않거나 기존의 데이터 증강 알고리즘을 사용하여 훈련한 경우 보다 높은 음질 및 명료도 평가 지수를 얻을 수 있었다. 각 평가 지표별 음성향상 전후의 점수를 비교한 경우 PESQ의 경우 평균 27.4%, STOI의 경우 5.2% 향상되어 데이터 증강 알고리즘 사용하지 않거나 기존의 데이터 증강 알고리즘을 사용하여 훈련한 경우 보다 높은 점수 향상율을 보여주었다.

destroyer engine을 포함한 일부 잡음의 경우 제안한 데이터 증강 알고리즘 보다 속도와 템포 변경 기반의 데이터 증강 알고리즘을 사용한 경우에 더 높은 음성 향상 성능을 보여 주었다. 주기성 잡음(periodic noise)의 경우 주파수 영역에서의 데이터 증강 알고리즘 보다는 시간 영역에서의 데이터 증강 알고리즘이 더 효과적인 것으로 추정된다. 잡음 특성별로 최적화된 데이터 증강 방법에 대한 연구가 진행된다면 추가적인 성능 향상이 가능할 것으로 예상된다.

본 연구의 가장 큰 의의는 많은 시간과 비용이 소모되는 데이터의 수집이나 추가적인 컴퓨팅 파워의 투입 없이도 제안한 데이터 증강 알고리즘만으로 음성 향상의 성능을 높일 수 있었다는 것이다. 본 연구가 적은 훈련 데이터로도 빠른 훈련과 최상의 성능을 보장해야 하는 인공지능 보청기나 음성인식 스피커 등의 분야에서 좋은 역량을 발휘 할 것으로 기대된다.

REFERENCE

[1] S. Boll, “Suppression of Acoustic Noise in Speech Using Spectral Subtraction,” *IEEE Transactions on Acoustic, Speech, and Signal Processomg*, Vol. ASSP-27, No. 2, pp. 113-120, 1979.

[2] P. Scalart and J.V. Filho, “Speech Enhancement Based on a Priori Signal to Noise Esti-mation,” *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 629-632, 1996.

[3] Y. Ephraim and H.L. Van Trees, “A Signal Subspace Approach for Speech Enhancement,” *Proceedings of 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 355-358, 1993.

[4] K.K. Paliwal and A. Basu, “A Speech Enhancement Method Based on Kalman Filtering,” *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 177-180, 1987.

[5] Y. Xu, J. Du, L. Dai, and C. Lee, “An Experimental Study on Speech Enhancement Based on Deep Neural Networks,” *IEEE Signal Processing Letters*, Vol. 21, No. 1, pp. 65-68, 2014.

[6] ITU-T P.862, *Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-End Speech Quality Assessment of Narrow-band Telephone Networks and Speech Codecs*, 2001.

[7] C.H. Taal, R.C. Hendrilks, R. Heusdens, and J. Jensen, “An Algorithm for Intelligibility Prediction of Time Frequency Weighted Noisy Speech,” *IEEE Transaction on Audio, Speech, and Language Processing,* Vol. 19, No. 7, pp. 2125-2136, 2011.

[8] T. Tran, J. Park, O. Kwon, K. Moon, S. Lee, K. Kwon, et al., "Classification of Leukemia Disease in Peripheral Blood Cell Images Using Convolutional Neural Network," *Journal of Korea Multimedia Society,* Vol. 21, No. 10, pp. 1150-1161, 2018.

[9] N. Jaitly and G.E. Hinton, "Vocal Tract Length Perturbation (VTLP) Improves Speech Recognition," *Proceedings of International Conference on Machine Learning Workshop on Deep Learning for Audio, Speech and Language*, pp. 925-660, 2013.

[10] J. Salamon and J.P. Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," *IEEE Signal Processing Letters*, Vol. 24, No. 3, pp. 279-283, 2017.

[11] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio Augmentation for Speech Recognition," *Proceeding of Sixteenth Annual Conference of the International Speech Communication Association*, pp3586-3589, 2015.

[12] L.J. Raphael, G.J. Borden, and K.S. Harris, *Speech Science Primer: Physiology, Acoustics, and Perception of Speech: Sixth Edition*, Lippincott Williams and Wilkins, Philadelphia, United States, 2012.

[13] D. Maurer, *Acoustics of the Vowel-Preliminaries,* Peter Lang AG, International Academic Publishers, Bern, Switzerland, 2016.

[14] J. Kim and M. Hahn, "Speech Enhancement Using a Two-Stage Network for an Efficient Boosting Strategy," *IEEE Signal Processing Letters*, Vol. 26, No. 5, pp. 770-774, 2019.

[15] A. Graves, A. Mohamed, and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," *Proceeding of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6645-6649, 2013.

[16] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, and D.S. Pallett, "DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM," *National Institute of Standards and Technology*, 1993.

[17] A. Varga and H.J.M. Steeneken, "Assessment for Automatic Speech Recognition II: Noisex-92: A Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition Systems," *Speech Communication*, Vol. 12, No. 3, pp. 247-251, 1993.

이 승 관

2014년 인천재능대학교 정보통신과 전문학사 졸업
2016년 한양사이버대학교 정보통신공학과 학사 졸업
2017년~현재 인하대학교 전자공학과 석사과정

관심분야 : Deep learning, Database, Speech Signal Processing

이 상 민

1987년 인하대학교 전자공학과 학사 졸업.
1989년 인하대학교 전자공학과 석사 졸업.
2000년 인하대학교 전자공학과 박사 졸업.

2006년~현재 인하대학교 전자공학과 교수
관심분야 : Bio-Signal Processing, Psyco-Acoustic, Brain-Machine Interface