ISSN : 1226-7244 (Print)
ISSN : 2288-243X (Online)
논문번호 19-04-35

j.inst.Korean.electr.electron.eng.Vol.23,No.4,1360~1365,December 2019
http://dx.doi.org/10.7471/ikeee.2019.23.4.1360
245

깊은 신경망 기반 음원 추적 기법

Sound Source Localization Method Based on Deep Neural Network

박 희 문*, 정 종 대*★

Hee-Mun Park*, Jong-Dae Jung*★

Abstract

In this paper, we describe a sound source localization(SSL) system which can be applied to mobile robot and automatic control systems. Usually the SSL method finds the Interaural Time Difference, the Interaural Level Difference, and uses the geometrical principle of microphone array. But here we proposed another approach based on the deep neural network to obtain the horizontal directional angle(azimuth) of the sound source. We pick up the sound source signals from the two microphones attached symmetrically on both sides of the robot to imitate the human ears. Here, we use difference of spectral distributions of sounds obtained from two microphones to train the network. We train the network with the data obtained at the multiples of 10 degrees and test with several data obtained at the random degrees. The result shows quite promising validity of our approach.

요 약

본 논문은 모바일 로봇과 자동제어 시스템에 적용될 수 있는 음원 위치 추적 시스템(Sound Source Localization, SSL)을 보여준다. 대부분 SSL의 기법은 음원 도달 시간차(Interaural Time Difference, ITD)와 음압 레벨의 차이(Interaural Level Difference, ILD)를 구하고, 마이크로폰 배열의 기하학적 원리를 이용하여 위치를 찾게 된다. 하지만 본 논문에서는 음원의 수평 각도를 구하기 위해 깊은 인공 신경망을 기반으로 한 다른 접근법은 제안한다. 인간의 귀를 모방한 로봇의 양쪽 마이크로폰에서 음원의 신호를 채집하여 연구에 사용했다. Network를 학습시키기 위해 양쪽 마이크로폰에서 얻어진 음원의 스펙트럼 분포 차이를 이용하였다. 각 10도 마다 채집한 데이터로 네트워크를 학습시켰고 임의의 각도에서 얻어진 데이터로 결과를 확인했다. 실험 결과 제안한 SSL의 접근 방식은 상당히 가능성이 있는 결과를 보여주었다.

Key words : Sound Source Localization(SSL), Deep Neural Network, Mobile Robot, Spectral Distribution, ITD, ILD

Ⅰ. 서론

인간과 같이 생각하고 행동하는 로봇의 연구가 활발하게 이루어지면서 로봇과 인간의 의사소통이 중요한 과제가 되었다. 로봇과 인간의 상호작용에서 기본적으로 중요한 것은 로봇이 명령을 내리는 사람의 위치를 정확하게 추정하는 것인데 이것을 음원 위치 추적(Sound Source Localization, SSL)

*School of Electrical, Electronics and Communication Engineering, Korea University of Technolgy and Education
★ Corresponding author
E-mail : jungjd@koreatech.ac.kr, Tel : +82-41-560-1164
※ Acknowledgment
Manuscript received Dec. 13, 2019; revised Dec. 20, 2019; accepted Dec. 26, 2019.

이라고 한다. 통상 SSL의 방법으로는 마이크로폰 배열로 음성 신호를 채집하여 신호가 도달한 시간차(Interaural Time Difference, ITD)와 음압차(Interaural Level Difference, ILD)를 이용하는 방법이 있다[1]. 그 중 ITD를 이용한 방법이 널리 사용되는데, 이 방법은 음원을 얼마나 정밀하게 수집하는지에 따라 성능이 좌우되고, 또한 주변 환경에 대해서도 취약한 단점이 있다[1].

본 논문에서는 사람이 소리를 듣고 판단하는 과정을 모방하여 두개의 마이크로폰을 가지고 신호를 채집했고, 사람의 청각시스템을 모사하기 위해 각각의 마이크로폰에 귓바퀴의 역할을 하는 판을 부착하였다. 이렇게 함으로써 같은 음원에 대해 양쪽 마이크로폰으로부터 약간의 상이한 주파수 스펙트럼을 얻을 수 있었는데 본 논문에서는 이 데이터와 인공신경망을 이용한 새로운 방식의 음원 추적 기법을 제안한다.

깊은 신경망(Deep Neural Network)은 구현하고자 하는 목적과 용도에 따라 합성곱 신경망(Convolution Neural Network, CNN), 장단기 메모리(Long Short Term Memoty, LSTM) 등 여러 네트워크가 개발되고 있으며 사용하는 환경에 맞추어 적절한 신경망의 구조를 선택하거나 설계하는 것은 매우 중요하다[2].

본 논문에서는 CNN 또는 LSTM 보다 구조가 간단한 DNN을 이용하여 양질의 성능을 발휘하는 것에 초점을 두었다. 채집된 음성 데이터를 전처리 없이 실험 한 결과와 Fast Fourier Transform(FFT)을 사용한 전처리 후 실험 한 결과와 비교하였고 전처리 과정이 성능에 어떤 영향이 있는지를 검토하였다.

II. Sound Source Localization

1. 시스템 설계

모바일 로봇의 크기를 고려하여 마이크로폰 사이의 거리를 34cm로 설계하여 제작하였다. 음원은 마이크로폰으로부터 1.3m 떨어진 거리에서 -90°로부터 90°까지 10°간격으로 채집하였다. 그림 1은 음원의 채집방법을 보여주며 여기서 R과 L은 좌우 마이크로폰, θ는 음원 채집각도이다.

채집한 음원 데이터는 별도의 윈도우 함수를 사용하지 않고 수신된 음압이 미리 설정한 임계값 보다 커지는 시점부터 0.32초 동안 sampling 하여 저장하였으며, 저장된 데이터의 일부는 Training Set으로 나머지는 Validation Set으로 사용되었다. Sampling을 시작하는 임계값은 주변 환경의 영향을 고려하여 반복적인 채집 실험을 통해 설정하였다. 채집한 총 데이터 개수는 10°의 배수가 되는 위치 당 33개씩 총 627개(33개/위치×19위치)로 구성되어있다. 그림 2는 본 논문에 사용된 음원 추적 장치의 구성을 보여준다.

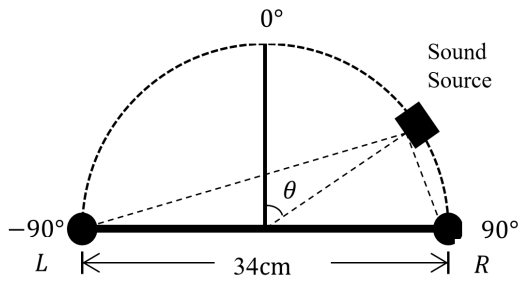


Fig. 1. Data Sampling Method.
그림 1. 음원 채집 방법

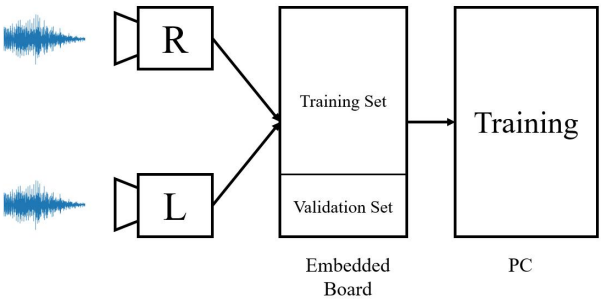


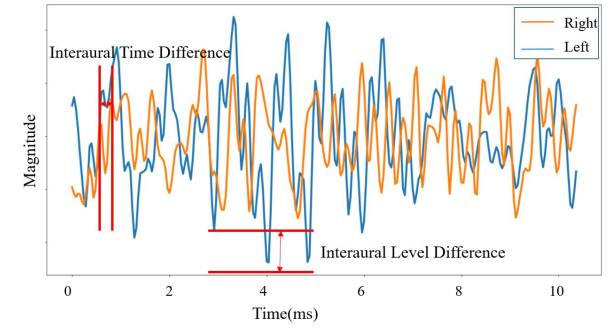
Fig. 2. System configuration.
그림 2. 시스템 구성

2. 데이터 전처리

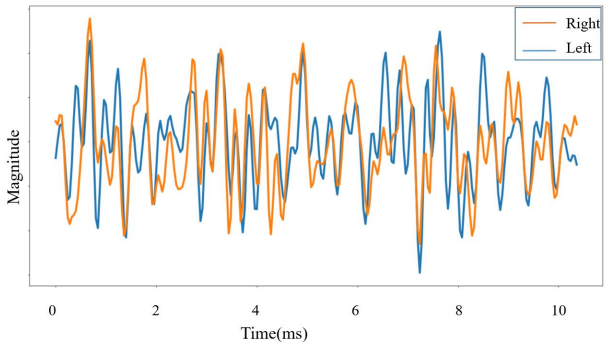
채집한 음원 데이터에서 ITD와 ILD는 그림 3과 같이 -90°와 90°에서 가장 크게 나타나며 0°에서 가장 작은 특성을 보인다.

그림 3을 보면 알 수 있듯이 ILD, ITD 데이터는 음원의 각도에 따른 변화가 균일하지 않아서 일관성 있는 추정 값을 얻기가 쉽지 않다. 같은 음원에 대해 주파수 변환을 하더라도 양쪽 마이크에서 얻어진 스펙트럼에는 차이가 있게 된다[4].

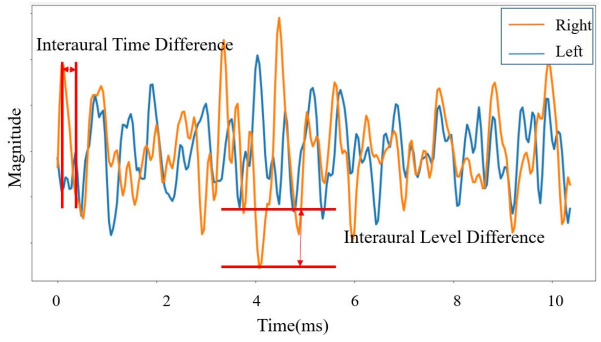
본 논문에서는 이 현상을 이용하여 음원을 추적하고자 하며, 이를 위해 주파수 변화를 하지 않은 신호와 주파수 변환으로 전처리를 한 신호를 이용하는 두 가지 실험을 실시하였다.



(a) -90° sound data



(b) 0° sound data



(c) 90° sound data

Fig. 3. (a) -90°, (b) 0°, (c) 90° Sound degree Data.
그림 3. (a) -90°, (b) 0°, (c) 90°에서 음원 데이터

신경망의 학습 효율과 성능을 높이기 위해서는 훈련에 사용되는 입력 데이터들 간의 상관관계가 매우 중요하다.

데이터 간의 상관관계를 비교하는 데는 피어슨 상관계수(Pearson Correlation Coefficients)가 널리 사용된다[3]. 피어슨 상관계수는 공분산 행렬(covariance matrix)을 이용하여 두 개 이상의 데이터가 어떤 상관관계를 가지고 분포하는지를 정량적으로 나타내는 것이다. 피어슨 계수(r)는 식(1)과 같이 구할 수 있다.

$$r = \frac{\sigma_{xy}}{\sigma_x \bullet \sigma_y}, \tag{1}$$

여기서, r은 피어슨 계수, σ_{xy} 는 x와 y의 공분산, σ_x는 x의 표준 편차, σ_y는 y의 표준 편차이다. 피어슨 계수 r의 범위는 $-1 \le r \le 1$ 이다. 1과 가까울수록 두 데이터는 정의 상관관계를 가지며 -1에 가까울수록 역의 상관관계 가지고 0이면 아무런 상관관계가 없음을 의미한다.[3] 그림 4는 90°에서 발생한 음원 데이터 중 임의로 10개의 Sample을 추출하여 Sample들의 상호 상관관계를 비교한 Matrix를 나타내며 횡축, 종축의 번호는 비교에 사용된 10개의 데이터 붙인 일련번호이다.

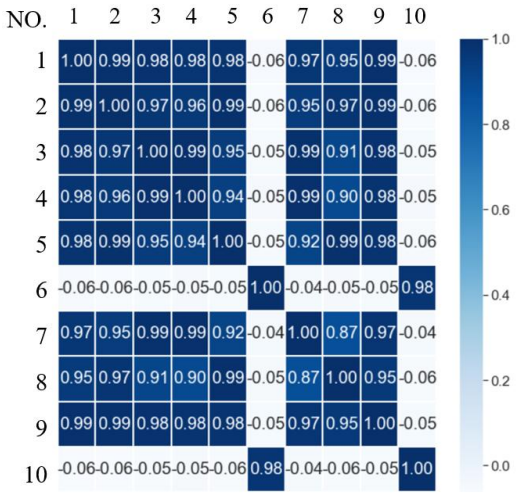


Fig. 4. Pearson Correlation Coefficients matrix of 90° data.
그림 4. 10개의 90° 데이터 간의 피어슨 상관계수 매트릭스

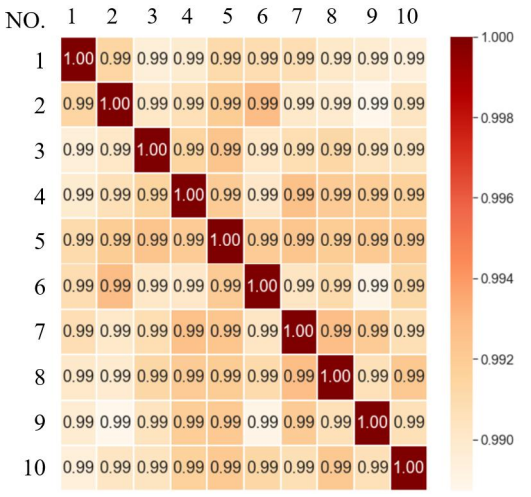


Fig. 5. Pearson Correlation Coefficient matrix of Preprocess 90° data.
그림 5. 전처리 후 10개의 90°데이터간의 피어슨 상관계수

그림 4의 피어슨 상관관계 Matrix를 보면 음원 데이터에 주파수 변환의 전처리를 수행하지 않은 경우 90°에서 얻어진 음원 데이터들 간에도 상관관

계가 없는 데이터들이 있다는 것을 알 수 있다. 반면, 그림 5는 전처리가 된 데이터들의 대한 상관관계를 보여주는데 Matrix에서 보듯이 같은 각도에서 얻어진 데이터들끼리는 일관성 있게 거의 1에 가까운 높은 상관관계를 보여주었다.

전처리 방법은 채집한 좌우 데이터에 FFT를 이용하여 주파수 스펙트럼을 구한 뒤 둘 사이의 차이(Difference)를 구하는 방식을 사용했다. 차이를 구한 이유는 음원과의 거리에 따른 음압정보를 이용하기 위함이며 식(2)와 같이 계산 된다.

$$\hat{X} = X_r - X_l \tag{2}$$

여기서, X_r, X_l은 각각 오른쪽 마이크로폰, 왼쪽 마이크폰에서 채집된 신호의 주파수 스펙트럼이며, $\hat{X}$은 전처리 결과 데이터로서 나중 신경망의 입력으로 사용되는 신호이다.

3. 깊은 인공 신경망(Deep Neural Network)

인공 신경망은 인간의 뇌기능을 모사한 시스템으로 뉴런들과 그 뉴런들을 연결 해주는 가중치들로 구성되는데, 그림 6의 (a)에 한 개의 은닉층(Hidden Layer)을 갖는 간단한 신경망 구조를 보여준다. 신경망에서 학습이란 신경망의 출력($\hat{y}$)과 학습 시키고자 하는 목표치(y)와의 오차가 최소화 되도록 반복적으로 가중치를 조절하는 것을 말한다[4]. 이와

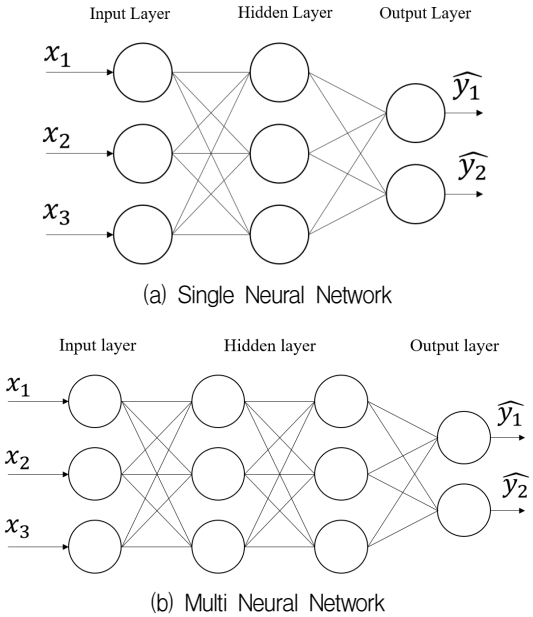


Fig. 6. Neural Network Structure.
그림 6. 인공신경망 구조

같은 신경망의 학습에서는 입력 x와 출력 $\hat{y}$사이의 상관관계를 알아야 하는데, 신경망이 단층(Single Layer)으로 구성 되어있는 경우는 이 상관관계를 바로 알 수 있지만, 그림 6의 (b)와 같이 다층(Multi Layer)구조인 경우는 그 관계를 직접적으로 구할 수가 없다.

이 문제를 해결하기 위한 방법으로 오차역전파(Error Back Propagation) 알고리즘이 사용되고 있으며 본 논문에서도 이 기법으로 신경망을 학습시켰다.[5] 일반적으로 하나의 은닉층을 갖는 신경망으로 대부분 임의의 함수를 모사할 수 있지만, 함수가 복잡해지면 은닉층에 사용되는 뉴런의 수가 급격히 증가 되어 신경망의 크기가 매우 커지고 또한 학습과 일반화(Generalization)가 제대로 일어나지 않을 수 있다는 문제가 있다[5].

따라서, 본 논문에서는 다양한 시뮬레이션 결과를 검토하여 3개의 은닉층을 갖는 DNN 구조를 사용하였고, 결과적으로 각 은닉층에 사용되는 뉴런 개수와 신경망 출력의 일반화 성능도 개선시킬 수 있었다.

III. 시뮬레이션 결과 및 분석

사용된 인공 신경망은 입력층과 3개의 은닉층 그리고 출력층으로 구성되어있다. 신경망에서 사용된 활성함수(Activation Function)들은 반복 된 시뮬레이션 결과를 검토하여 선택하였으며 각 층에 따라 Sigmoid 함수와 Relu(Rectified Linear Unit)함수를 적절하게 사용하였다. 총 학습 데이터는 608×2500의 행렬로 구성되며 608은 데이터의 개수이고 2500은 신경망에 입력되는 데이터의 길이를 나타낸다. 그림 7은 본 연구에서 사용된 DNN의 모델을 나타낸다.

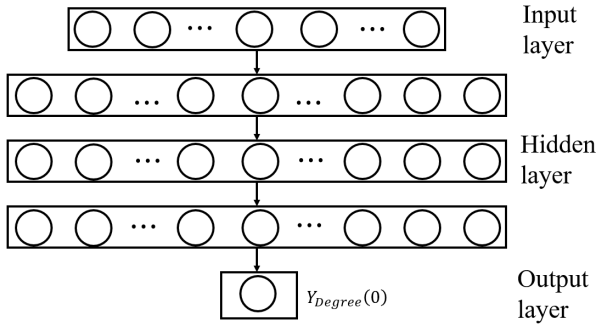


Fig. 7. DNN model implemented by tensorflow.
그림 7. 텐서플로우로 구현한 DNN 모델

신경망은 텐서플로우를 이용하여 구현하였고, GPU

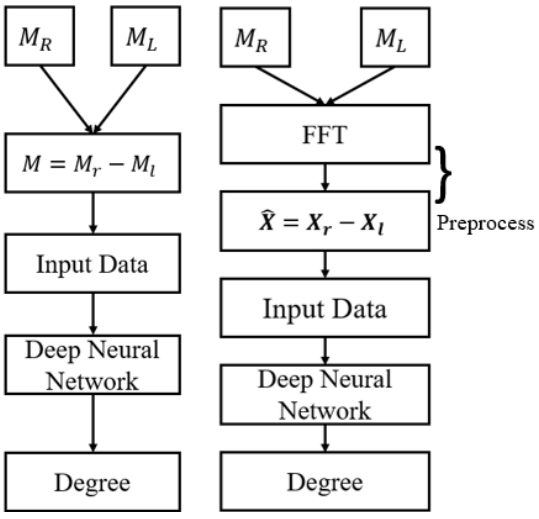
는 NVDIA Geforce GTX 1060 3GB를 사용하여 학습을 진행했다. 표 1은 학습에 사용된 음원 데이터의 구성을 나타낸다.

Table 1. Sound Data Configuration.
표 1. 음원 데이터 구성

	Without Preprocessing	With Preprocessing
Sampling Rate	16 khz	
Data point (Size of input Neural Network)	2500 point	5000 point
Sampling Degree	-90° ~ 90°	

본 논문에서 제안한 전처리 효과의 우수성을 검증하기 위하여, 전처리 하지 않은 입력 데이터와 전처리 후 입력 데이터를 사용하여 같은 구조의 신경망으로 학습 시켰으며 각각의 과정을 그림8에 보인다. 그림 8(a)는 전처리를 하지 않은 방식이며, (b)는 전처리를 한 방식이고 M_R과 M_L은 양쪽 마이크로폰에서 채집한 음원 데이터이다.

실제 신경망의 훈에는 우측신호-좌측신호를 입력으로 사용하였으며 이것은 그림 8에서 보듯이 전처리가 없는 경우는 채집된 마이크로폰 데이터 $M_R - M_L$이고, 전처리가 있는 경우에는 주파수 변환된 데이터 $X_r - X_l$이다.



(a) Without Preprocessing (b) With Preprocessing
Fig. 8. Process of SSL Method.
그림 8. 음원추적 기법의 과정

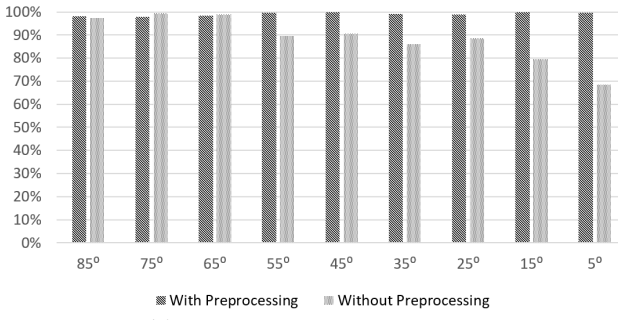
한편 학습된 신경망의 성능을 평가하기 위한 출력 값의 Error와 Accuracy는 식(3), 식(4)와 같이 정의 하였다.

$$error = \alpha_t - \alpha_o, \tag{3}$$

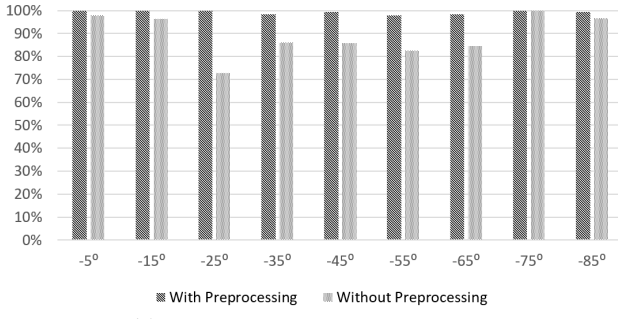
$$Accuracy = \frac{180 - error}{180}, \tag{4}$$

여기서, $error$는 출력 값 오차, α_t는 해당 입력의 목표 값 α_o는 해당 모델의 출력 값, $Accuracy$는 정확도이다.

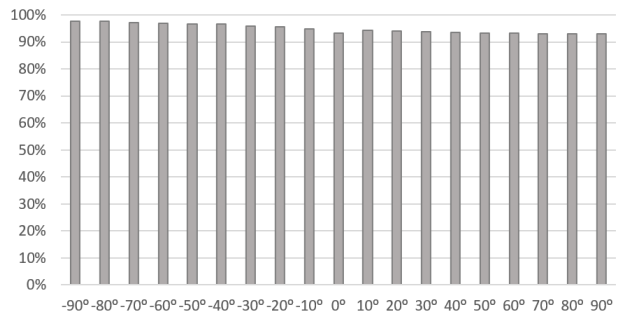
그림 9의 (a)와 (b)를 보면 훈련되지 않은 각도에서 채집된 모든 데이터에 대해서 FFT로 전처리를 한 경우가 그렇지 않은 경우보다 월등한 각도 추정 능력을 가지는 것을 알 수 있다. 또한, 그림 9에서와 같이 전처리를 하지 않은 모델의 정확도는 평균 89%이고, 전처리를 한 모델의 정확도는 평균 99%로서 정확도가 두드러지게 향상되었음을 알 수 있다.



(a) Result of between 85° and 5°



(b) Result of between -5° and -85°



(c) Result of CNN
Fig. 9. Result of Simulation.
그림 9. 실험 결과

그림 9의 (c)는 Deep Learning에 널리 사용되는

CNN(Convolutional Neural Network)을 이용하여 학습시킨 모델의 결과를 보여준다[6]. CNN은 본 논문에 사용된 DNN 보다 그 구조가 훨씬 복잡한데 그래프에서 보이듯이 CNN 모델의 평균 정확도는 96%로 전 각도영역에서 본 논문의 정확도가 3% 정도 우수한 성능을 보이는 것을 알 수 있다.

Ⅳ. 결론

본 논문에서는 음원추적에서 널리 사용되었던 ITD의 단점을 보완하기 위해 인공신경망(DNN)을 사용하였다. 또한 음원 추적의 성능을 더욱 개선하기 위해 수집된 음원을 FFT로 전처리하는 방식을 제안하였다. 실험은 10° 간격으로 수집된 음원으로 학습 시키고 학습된 신경망에 학습되지 않은 임의의 각도의 음원을 입력하여 그 위치를 추정하게 하는 방식으로 진행 하였다. 그림 9에서 보듯이 본 논문에서 제안한 DNN + 전처리 모델은 모든 각도에서 우수한 추정 성능을 보여주었다.

비록 본 논문의 결과가 제한적인 음원, 거리 및 sample의 개수에 대해 얻어진 것이지만 향후 다양한 음원의 거리와 종류 및 sample 개수에 대한 강인한 음원 추적 시스템을 개발하는데 소중한 기초자료가 될 수 있을 것이다.

References

[1] S. H. Oh and K. S. Park, “Optimal Acoustic Sound Localization System Based on a Tetrahedron-Shaped Microphone Array,” *Journal of KIISE,* Vol.43, No.1, pp.13-26, 2016.
DOI: 10.5626/JOK.2016.43.1.13

[2] Yalta, N. Nakadai, K, & Ogata, “Sound source localization using deep learning models,” *Journal of Robotics and Mechatronics*, Vol.29, No.1, pp.37-48, 2017. DOI: 10.20965/jrm.2017.p0037

[3] Wikipidia, “Correlation and dependence,” https://en.wikipedia.org/wiki/Correlation_and_dependence

[4] Wikipidia, “Sound localization,” https://en.wikipedia.org/wiki/Sound_localization

[5] Ian Goodfellow, Deep Learning, MIT PRESS, 2016.

[6] Suvorov. Dmitry and Dong. Ge and Zhukov. Roman, “Deep Residual Network for Sound Source Localization in the Time Domain,” *Journal of Engineering and Applied Sciences*, Vol.13, No.13, pp.5096-5104, 2018.
DOI: 10.3923/jeasci.2018.5096.5104

BIOGRAPHY

Hee-Mun Park (Member)

2017 : BS degree in Electrical, Electronics and Communication Engineering, Korea University of Technolgy and Education
2019 : MS degree in Electrical, Electronics and Communication Engineering, Korea University of Technolgy and Education

Jong-Dae Jung (Member)

1979 : BS degree in Electrical Engineering, Seoul National University.
1982 : MS degree in Electrconics Engineering, Seoul National University.
1990 : PhD degree in Electronics Engineering, Seoul National University.
1993~ : Professor, School of Electrical, Electronics and Communication Engineering, Korea University of Technolgy and Education