특집논문 (Special Paper)
방송공학회논문지 제24권 제2호, 2019년 3월 (JBE Vol. 24, No. 2, March 2019)
https://doi.org/10.5909/JBE.2019.24.2.227
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

가상현실 음향을 위한 심층신경망 기반 사운드 보간 기법

최 재 규[a)], 최 승 호[a)‡]

A Sound Interpolation Method Using Deep Neural Network for Virtual Reality Sound

Jaegyu Choi[a)] and Seung Ho Choi[a)‡]

요 약

본 논문은 가상현실 음향 구현을 위한 심층신경망 기반 사운드 보간 방법에 관한 것으로서, 이를 통해 두 지점에서 취득한 음향 신호들을 사용하여 두 지점 사이의 음향을 생성한다. 산술평균이나 기하평균 같은 통계적 방법으로 사운드 보간을 수행할 수 있지만 이는 실제 비선형 음향 특성을 반영하기에 미흡하다. 이러한 문제를 해결하기 위해서 본 연구에서는 두 지점과 목표 지점의 음향신호를 기반으로 심층신경망을 훈련하여 사운드 보간을 시도하였으며, 실험결과 통계적 방법에 비해 심층신경망 기반 사운드 보간 방법의 성능이 우수함을 보였다.

Abstract

In this paper, we propose a deep neural network-based sound interpolation method for realizing virtual reality sound. Through this method, sound between two points is generated by using acoustic signals obtained from two points. Sound interpolation can be performed by statistical methods such as arithmetic mean or geometric mean, but this is insufficient to reflect actual nonlinear acoustic characteristics. In order to solve this problem, in this study, the sound interpolation is performed by training the deep neural network based on the acoustic signals of the two points and the target point, and the experimental results show that the deep neural network-based sound interpolation method is superior to the statistical methods.

Keyword : VR sound, Deep Neural Network, Sound Interpolation

a) 서울과학기술대학교 전자IT미디어공학과(Dept. of Electronic and IT Media Engineering, Seoul National University of Science and Technology)
‡ Corresponding Author : 최승호(Seung Ho Choi)
E-mail: shchoi@seoultech.ac.kr
Tel: +82-2-970-6461
ORCID: https://orcid.org/0000-0001-8626-8355
※ 이 논문의 연구 결과 중 일부는 “2018년 한국방송·미디어공학회 추계학술대회”에서 발표한 바 있음.
※ This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2016-0-00144, Moving Free-viewpoint 360VR Immersive Media System Design and Component Technologies).
· Manuscript received January 8, 2019; Revised March 18, 2019; Accepted March 18, 2019.

Ⅰ. 서 론

본 연구는 UCC(User Create Contents)를 이용한 시청자 이동형 자유시점 360VR 실감미디어 제공과 관련된 연구이다. 현재 VR 미디어를 취득하기 위해서 주로 특수한 장비를 사용하며, 이러한 방법은 제한된 사용자만이 VR(Virtual Reality) 콘텐츠를 제작할 수 있다는 한계를 갖는다. 또한 현재 일반인들이 경험하는 대다수의 가상현실 콘텐츠는 촬영 지점만을 중심으로 360도 회전하여 시청자의 시점이 제한되는 한계가 존재한다. 이 한계를 극복하기 위해 최근 시청자 자유시점형 360VR 콘텐츠에 대한 연구가 활발히 진행 중이다. 시청자 이동형 실감 미디어를 제공하기 위해서는 시청자 이동에 따른 임의의 좌표에서 촬영한 데이터가 필요하다. 하지만 무수히 많은 지점에서 촬영한 데이터를 취득하기에는 현실적으로 한계가 있으므로 제한된 데이터를 이용해 임의의 지점에서 취득한 데이터를 생성 하는 기술이 필요하다. 특히 일반 사용자가 취득한 UCC를 이용해 가상현실 음향을 생성하기 위해선 가상의 지점에서의 음향 신호를 주어진 데이터만으로 생성할 수 있어야 하고, 이를 위해선 사운드 보간(sound interpolation) 기법이 필요하다. 통계적 방법을 이용해 사운드 보간을 진행할 수 있으나 실제 음향환경의 비선형 특성을 잘 반영할 수 없다. 이러한 문제를 해결하기 위하여 본 논문에서는 심층신경망(deep neural network, DNN)을 이용한 사운드 보간 기법을 제안한다. 또한, 임의의 방향각에 해당하는 HRIR(head related impulse response)을 이용하면 360도 전방위 음향을 생성할 수 있으나, 실제 음향 환경에서는 이러한 HRIR을 얻을 수 없다. 본 연구의 향후 최종 목표는 시점이 다른 두 지점에서 촬영한 데이터를 이용하여 취득하지 못한 임의의 시점의 사운드를 추정하는 것이며, 본 논문은 DNN 기반 사운드 보간 기법을 통해 취득하지 못한 지점의 사운드를 추정하기 위한 사전 연구이다.

Ⅱ. 심층신경망을 이용한 사운드 보간 기법

본 연구에서는 실제 음향환경의 비선형 특성을 반영하여 성능을 개선하기 위해 심층신경망을 이용한 사운드 보간 방법을 개발하였다. 두 지점에서 받은 음향 신호의 단구간 스펙트럼(short-time spectrum)을 구한 뒤 이 두 지점의 데

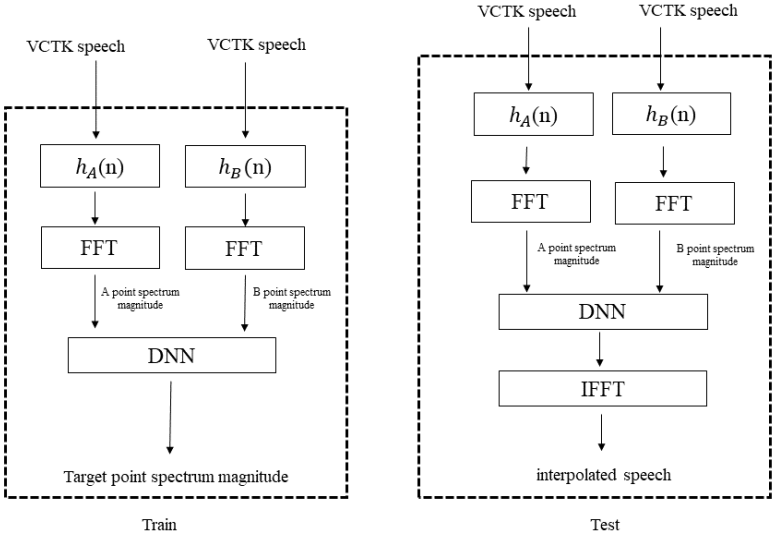

그림 1. 시스템 흐름도
Fig. 1. System flowchart

이터를 입력으로 사용하고 목표 지점의 단구간 스펙트럼이 출력되도록 심층신경망을 훈련시킨다. 이러한 과정을 통해서 제한된 음향 데이터만을 이용해 두 지점 사이의 취득하지 못한 좌표 지점의 음향 데이터를 얻어낼 수 있다. 그림 1은 시스템의 흐름도이다. 먼저 VCTK 음성[1]을 A지점 및 B지점의 HRIR(head related impulse response) 혹은 RIR (room impulse response)과 콘볼루션하여 방향성을 가진 음성 데이터를 생성한다. 이후 FFT(fast Fourier transform)를 통하여 스펙트럼을 얻어내어 두 지점의 스펙트럼 크기를 입력으로 구성한 후 목표 지점의 스펙트럼 크기가 출력되도록 훈련한다. 이후 앞선 훈련 과정에 겹치지 않는 데이터를 이용해 두 지점의 스펙트럼 크기를 입력했을 때 목표 지점의 스펙트럼 크기가 출력되는지 테스트하는 과정을 거친다. 테스트 과정을 거쳐 얻은 목표 지점의 스펙트럼 크기와 정면의 위상을 이용해 IFFT(inverse fast Fourier transform)하여 음향 신호를 추정하였다.

본 연구에서의 심층신경망 구조는 그림 2와 같으며, A 지점과 B 지점에서 취득한 음향신호의 스펙트럼 크기 $\{|X_A(0)|,\cdots,|X_A(N)|\}$와 $\{|X_B(0)|,\cdots,|X_B(N)|\}$를 입력으로 하고 두 지점 사이의 가상의 지점에서의 스펙트럼 크기를 $\{|X_{ref}(0)|,\cdots,|X_{ref}(N)|\}$ 출력으로 한다. 그림 2의 입력과 출력은 스펙트럼의 크기를 예시로 나타낸 것이다.

Ⅲ. 실험 및 결과

1. 실험 내용

실험에 사용한 심층신경망의 은닉층 개수는 3이며 입력 노드는 A와 B지점의 각각 frequency bin 257개를 더한 514개이고 출력노드는 목표 지점인 C에 해당하는 frequency bin 257개로 구성했다. 이때 frequency bin은 프레임별 512-point FFT ($\{X(k),k=0,1,\cdots,511\}$)로부터 얻은 half spectrum magnitude($\{|X(k)|,k=1,\cdots,256\}$)와 1개의 DC성분($|X(0)|$)을 포함한 것이다. 활성함수 (activation function)로는 노드의 입력 x에 노드 출력이 max(0, x)인 ReLU (Rectified Linear Units)[2]를 사용하였으며 활성함수 이후에 drop out[3]도 사용하였다. 또한 학습시 최적화 알고리즘은 ADAM(ADAptive Moment estimation)[4]을 사용하였다. 이렇게 훈련된 심층신경망을 통해 목표 지점에 해당하는 스펙트럼 크기를 추정한다.

2. 머리전달함수 기반 합성음원에 대한 실험

실험에 사용한 데이터는 VCTK 음성과 PKU-IOA HRTF 데이터베이스[5]이다. 머리전달함수(head related transfer

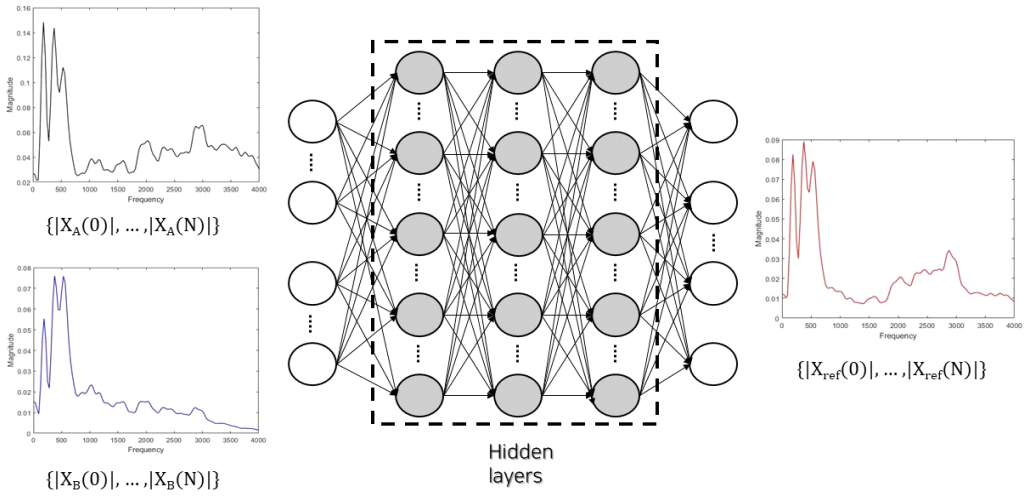

그림 2. 심층신경망 구성도
Fig. 2. Structure of deep neural network

function, HRTF)는 정면을 0도 오른쪽이 90도 왼쪽을 270도로 정의하며 2채널 사운드이다. 본 실험은 우선 음성신호에 대한 것이므로 16,000 Hz로 다운샘플링하여 실험하였다. 모노 사운드인 VCTK 데이터와 각각 0도와 45도 그리고 90도에 해당하는 HRIR(head related impulse response)을 콘볼루션하여 스테레오 음향 신호를 생성하였다.

그림 3은 통계적 방법과 기준 신호 그리고 심층신경망을 이용해 구한 스펙트럼의 크기를 비교한 것으로서, 0에서 4,000 Hz에 해당하는 부분을 확대하여 도시한 것이다. 통계적 방법은 각각 양쪽 지점의 스펙트럼 크기를 산술평균과 기하평균으로 보간하는 방법을 의미하며 범례에서 각각 Arithmetic과 Geometric에 해당한다. 그림에서 알 수 있듯이 45도에 해당하는 스펙트럼과 0도와 90도 지점 스펙트럼의 산술 평균과 기하평균을 이용해 구한 예상 값이 차이가 나는 것을 확인할 수 있다. 이에 반해 심층신경망(범례의 DNN)을 이용해 얻은 추정치는 실제 값과 거의 일치함을 알 수 있다. 객관적 성능 비교를 위해 선형 크기 스펙트럼(linear magnitude spectrum)의 RMSE(root mean square error)를 사용하였다. 표 1은 머리전달함수 기반 합성음원에 대한 사운드 보간 기법의 음성 데이터 RMSE 결과이다.

표 1의 0/45/90은 0도와 90도의 스펙트럼 크기를 이용해 45도의 추정 스펙트럼 크기를 얻어냄을 의미한다. 표 1의 결과를 통해 산술 및 기하 평균으로 구한 추정치와 비교했을 때 심층신경망을 이용하여 구한 추정치의 RMSE가 큰 폭으로 감소한 것을 확인할 수 있다. 특히 90도와 270도의 스펙트럼 크기를 이용해 0도인 정면의 스펙트럼 크기를 추정한 결과 값이 가장 큰 값으로 감소한 것을 확인할 수 있다. 이러한 결과를 통해서 각도가 변함에도 불구하고 심층신경망을 기반으로 추정한 값이 가장 기준 값과 일치한다는 것을 확인할 수 있었다.

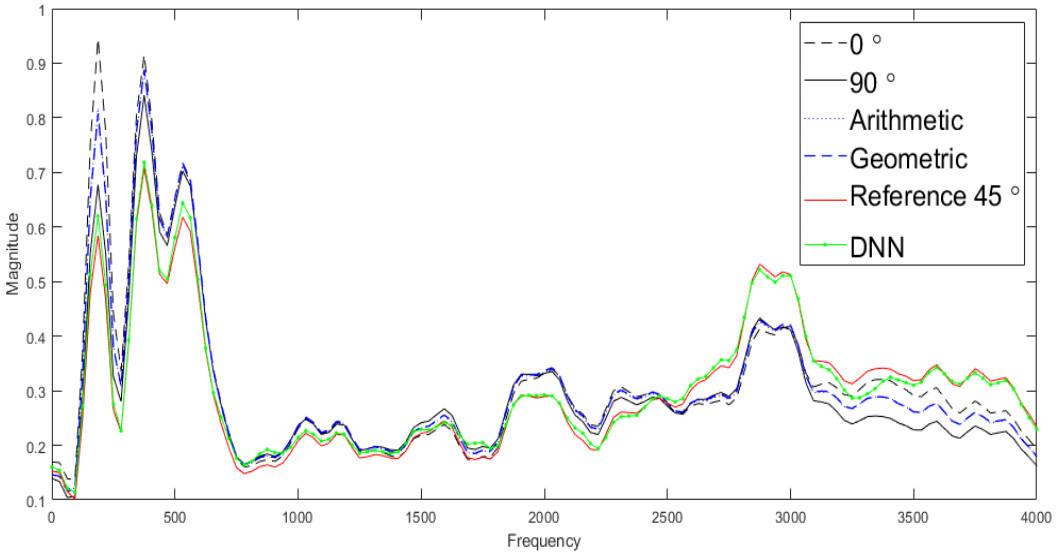

그림 3. 머리전달함수 합성 음원에 대한 스펙트럼 예시
Fig. 3. Spectrum example based on the sound generated by head-related transfer function

표 1. 머리전달함수 기반 합성음원에 대한 사운드 보간 기법의 음성 데이터 RMSE 결과
Table 1. RMSE result of speech data based on the sound generated by head-related transfer function

RMSE / Method	Speech					
	Left			Right		
	0/45/90 (degree)	0/270/315 (degree)	90/0/270 (degree)	0/45/90 (degree)	0/270/315 (degree)	90/0/270 (degree)
Arithmetic Mean	0.39	0.31	0.59	0.18	0.59	0.61
Geometric Mean	0.40	0.37	0.35	0.18	0.48	0.40
DNN	0.36	0.19	0.14	0.13	0.33	0.14

표 2. 머리전달함수 기반 합성음원에 대한 사운드 보간 기법의 스펙트럼 RMSE 결과
Table 2. RMSE result of spectrum magnitude based on the sound generated by head-related transfer function

RMSE / Method	Spectrum					
	Left			Right		
	0/45/90 (degree)	0/270/315 (degree)	90/0/270 (degree)	0/45/90 (degree)	0/270/315 (degree)	90/0/270 (degree)
Arithmetic Mean	0.23	0.29	0.6	0.18	0.48	0.61
Geometric Mean	0.26	0.36	0.36	0.18	0.37	0.42
DNN	0.12	0.12	0.12	0.10	0.09	0.11

표 2는 각도별로 추정한 스펙트럼 크기와 목표 값과의 RMSE 값을 나타낸 표이다. 표 1과 같이 모든 각도에서 DNN으로 추정한 값과 기준 값과의 RMSE 값이 가장 작음을 확인 할 수 있었고 이는 통계적 방법보다 DNN을 통해 얻어낸 결과 값이 통계적 방법보다 우수하다는 것을 확인할 수 있었다.

3. 잔향 환경 합성음원에 대한 실험

잔향 환경에서 유효함을 확인하기 위해 sound beamforming 데이터를 이용해 실험해 보았다. 이를 위해 MARDY 데이터베이스의 실내 임펄스응답(room impulse response)[6]을 활용하였다. MARDY 데이터베이스는 8개의 마이크를 일렬로 배치한 후 특정 지점의 스피커에서 나는 소리를 녹음한 것이다. MARDY 데이터 베이스의 샘플링 주파수는 VCTK 데이터베이스와 동일한 48 kHz이다. 이를 16,000 Hz로 다운샘플링 후 VCTK 데이터와의 콘볼루션을 통해 잔향환경 음원을 생성하였다. 그림 4와 같은 배치에서 L-speaker의 출력 음향을 Mic1과 Mic5로 취득한 후 이를 통해 Mic3의 음향 신호를 생성하는 실험을 진행했다.

앞선 실험과 동일한 방법으로 단구간 스펙트럼을 구하였다. 입력 노드는 Mic1의 frequency bin 257차와 Mic5의 frequency bin 257차를 합한 514개로 구성했고, 출력 노드는 Mic3의 frequency bin인 257차로 훈련을 진행했다. 이후 과정은 전 실험과 동일하게 진행하였다.

그림 5는 통계적 방법과 기준 지점의 스펙트럼 그리고 심층신경망을 이용해 구한 음향 신호의 스펙트럼의 크기를 비교한 그래프이다. 비교를 위해 차이를 주로 보이는 0에서 1,000 Hz에 해당하는 그래프를 확대 도시한 것이다. 그림에서 알 수 있듯이 통계적 방법으로 구한 스펙트럼보다 심

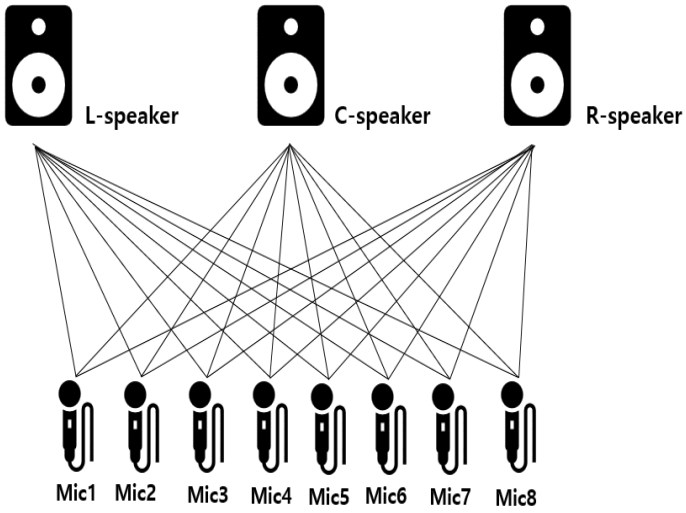

그림 4. 스피커와 마이크의 배치 [6]
FIg. 4. Array of speaker and microphone

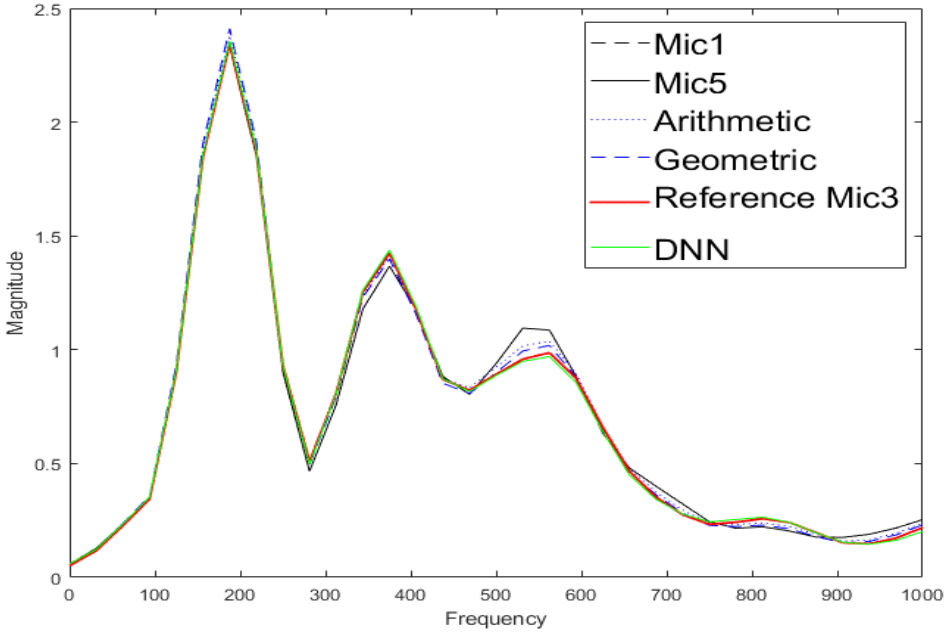

그림 5. 잔향 환경의 음원에 대한 사운드 보간 기법의 스펙트럼 예시
Fig. 5. Spectrum example based on the sound generated by room impulse response

층신경망을 통해 얻은 스펙트럼이 기준 신호와 더 유사함을 확인할 수 있었다.

표 3. 잔향 환경 음원에 대한 사운드 보간 기법의 RMSE 결과
Table 3. RMSE result based on the sound generated by room impulse response

RMSE / Method	Speech	Spectrum magnitude
Arithmetic Mean	0.16	0.12
Geometric Mean	0.16	0.13
DNN	0.09	0.08

표 3은 객관적 수치 비교를 위해 기준 되는 신호와 RMSE를 구한 것이다. 표를 통해 심층신경망을 이용해 구한 추정치의 RMSE가 가장 작음을 확인할 수 있다.

Ⅳ. 결 론

본 논문에서는 UCC를 이용한 시청자 이동형 가상현실 음향 구현에 있어서 두 지점 사이의 음향 신호를 생성할 경우, 심층신경망을 이용한 사운드 보간법에 대한 알고리즘 및 실험 결과를 다루었다. 우선, 두 지점에서 받은 음향 신호의 단구간 스펙트럼으로부터 통계적 방법인 산술평균과 기하평균을 이용해 사운드 보간을 진행해보았으나 실제 비선형 음향 특성을 반영하지 못해 성능이 미흡함을 확인하였다. 이를 개선하기 위해 심층신경망을 이용한 사운드 보간 기법을 개발하였다. 머리전달함수 기반 실험과 잔향환경 실험을 통해 심층신경망 기반 사운드 보간 기법이 통계적 방법에 비해 성능이 우수함을 확인할 수 있었다. 향후 본 연구를 기반으로 전방위 360도 임의의 지점의 일반 음향 사운드를 보간하는 연구를 진행할 계획이다. 또한 객관적인 평가뿐만 아니라 주관적인 음질 평가를 진행할 계획이다.

참 고 문 헌 (References)

[1] Veaux Christophe, Yamagishi Junichi, and MacDonald Kirsten, ”CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit,” *The Centre for Speech Technology Research (CSTR)*, 2016.

[2] V. Nair and G. E. Hinton, “Rectified linear units improve restricted Boltzmann machines,” in *Proc. 27th Int. Conf. Machine Learning,* pp. 807-814, 2010.
[3] Vu Pham, Théodore Bluche, Christopher Kermorvant, and Jérôme Louradour, “Dropout improves recurrent neural networks for handwriting recognition,” *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference*, pp. 285‑290, IEEE, 2014.
[4] D. P. Kingma and J. L. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
[5] T. Qu, Z. Xiao, M. Gong, Y. Huang, X. Li, and X. Wu, “Distance dependent head-related transfer functions measured with high spatial resolution using a spark gap,” *IEEE Trans. on Audio, Speech and Language Processing,* vol. 17, no. 6, pp. 1124-1132, 2009.
[6] J. Wen, N. Gaubitch, E. Habets, T. Myatt, P. Naylor, “Evaluation of speech dereverberation algorithms using the MARDY database”, *Proc. Int. Workshop Acoust. Echo Noise Control,* pp. 1-4, 2006.

저 자 소 개

최 재 규

- 2013년 - 현재 : 서울과학기술대학교 전자IT미디어공학과 학사과정
- ORCID : https://orcid.org/0000-0001-9354-5971
- 주관심분야 : 신호처리, 딥러닝, 입체음향

최 승 호

- 1987년 - 1991년 : 한양대학교 전자공학과 학사
- 1991년 - 1993년 : 한국과학기술원 전기 및 전자공학과 석사
- 1993년 - 1999년 : 한국과학기술원 전기 및 전자공학과 박사
- 1995년 - 2002년 : 삼성종합기술원 전문연구원
- 2002년 - 현재 : 서울과학기술대학교 전자IT미디어공학과 교수
- ORCID : https://orcid.org/0000-0001-8626-8355
- 주관심분야 : 음성신호처리, 음성인식, 화자인식, 음성코딩, 가상현실 음향