가상현실 음향을 위한 심층신경망 기반 사운드 보간 기법

*최재규 **최승호
서울과학기술대학교
*gyuuu_@naver.com **shchoi@seoultech.ac.kr

A Sound Interpolation Method Based on Deep Neural Networks for Virtual Reality Sound

*Jaegyu Choi **Seung Ho Choi
Seoul National University of Science and Technology

요약

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 지점들과 목표 지점의 음향신호를 기반으로 심층신경망을 훈련하여 사운드 보간을 시도하였으며, 실험결과 통계적 방법에 비해 심층신경망 기반 사운드 보간 방법의 성능이 우수함을 보였다.

1. 서론

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. 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.

2. 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 데이터를 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. 본 연구에서의 심층신경망 구조는 그림 1과 같으며, A 지점과 B 지점에서 취득한 음향신호의 스펙트럼 크기 $\{|X_A(1)|,\cdots,|X_A(N)|\}$와 $\{|X_B(1)|,\cdots,|X_B(N)|\}$를 입력으로 하고 두 지점 사이의 가상의 지점에서의 스펙트럼 크기를 $\{|X_{ref}(1)|,\cdots,|X_{ref}(N)|\}$ 출력으로 한다.

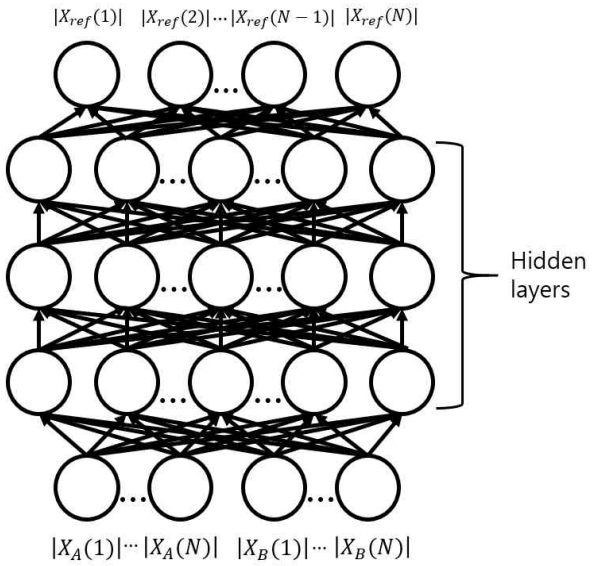

그림 1. 심층신경망의 구조

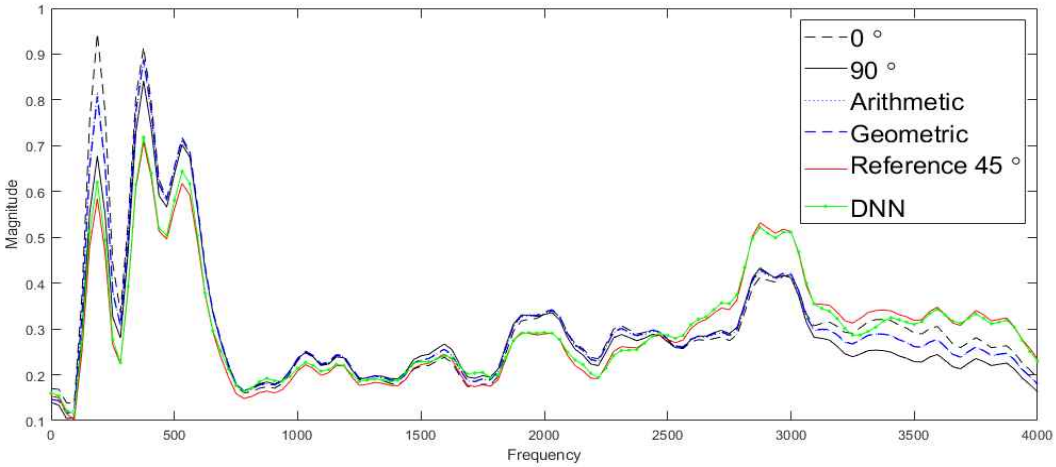

그림 2. 머리전달함수 합성 음원에 대한 스펙트럼 예시

3. 실험 및 결과

실험에 사용한 심층신경망의 은닉층 개수는 3이며 입력 노드는 A와 B지점의 각각 주파수 해상도(Frequency bin) 257개를 더한 514개이고 출력노드는 목표 지점인 45도에 해당하는 주파수 해상도 257개로 구성했다. 활성함수 (activation function)로는 노드의 입력 x에 노드 출력이 max(0, x)인 ReLU(Rectified Linear Units) [1]를 사용하였으며 활성함수 이후에 drop out [2]도 사용하였다. 또한 학습시 최적화 알고리즘은 ADAM(ADAptive Moment estimation) [3]을 사용하였다. 이렇게 훈련된 심층신경망을 통해 목표 지점에 해당하는 스펙트럼 크기를 추정한다.

3.1 머리전달함수 기반 합성음원에 대한 실험

실험에 사용한 데이터는 VCTK 음성 [4]과 PKU-IOA HRTF 데이터베이스 [5]이다. 머리전달함수(head related transfer function, HRTF)는 정면을 0도 오른쪽이 90도 왼쪽을 270도로 정의한다. 실험을 위해 두 데이터는 16,000 Hz로 다운샘플링 하였으며, 모노 사운드인 VCTK 데이터와 각각 0도와 45도 그리고 90도에 해당하는 HRIR(head related impulse response)을 콘볼루션하여 스테레오 음향 신호를 생성했다. 그림 2는 통계적 방법과 기준 신호 그리고 심층신경망을 이용해 구한 스펙트럼의 크기를 비교한 것으로서, 0에서 4,000 Hz에 해당하는 부분을 확대하여 도시한 것이다. 범례의 Arithmetic과 Geometric은 각각 산술 평균과 기하 평균으로 얻은 스펙트럼의 크기를 의미한다. 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 RMSE(Root Mean Square Error)를 사용하였다. 표 1은 머리전달함수 기반 합성음원에 대한 사운드 보간 기법의 RMSE 결과이다.

표 1. 머리전달함수 기반 합성음원에 대한 사운드 보간 기법의 RMSE 결과

방법	RMSE			
	왼쪽		오른쪽	
	파형	스펙트럼 크기	파형	스펙트럼 크기
산술평균	0.3919	0.2292	0.1754	0.1790
기하평균	0.4020	0.2560	0.1758	0.1793
DNN	0.3596	0.1175	0.1312	0.1011

산술 및 기하 평균으로 구한 추정치와 비교했을 때 심층신경망을 이용하여 구한 추정치의 RMSE가 큰 폭으로 감소한 것을 확인할 수 있다.

3.2 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 활용하였으며, 이를 16,000 Hz로 다운샘플링 후 VCTK 데이터와의 콘볼루션을 통해 잔향환경 음원을 생성하였다. 그림 3과 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 주파수 해상도 257차와 Mic5의 주파수 해상도 257차를 합한 514개로 구성했고, 출력 노드는 Mic3의 주파수 해상도인 257차로 훈련을 진행했다. 이후 과정은 전 실험과 동일하게 진행하였다.

그림 4는 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 스펙트

럼보다 심층신경망을 통해 얻은 스펙트럼이 기준 신호와 더 유사함을 확인할 수 있다.

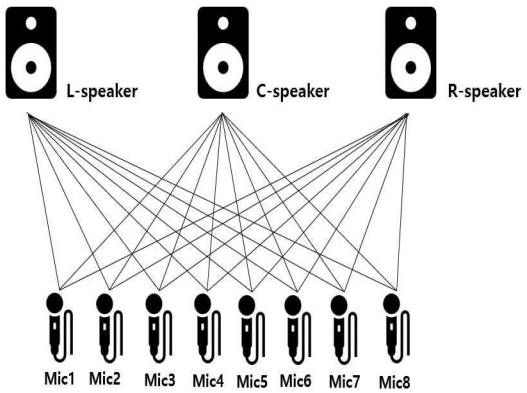

그림 3. 스피커와 마이크의 배치 [6]

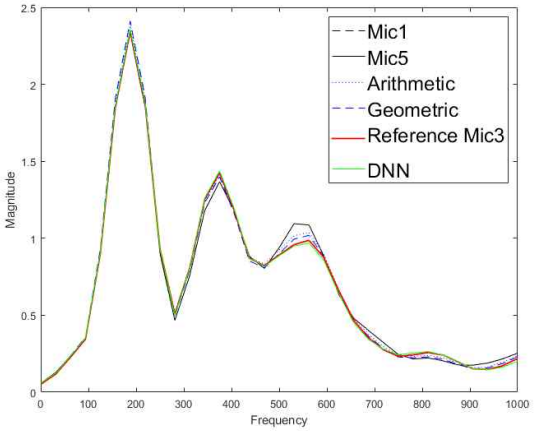

그림 4. 잔향 환경의 음원에 대한 사운드 보간 기법의 스펙트럼 예시

표 2는 객관적 수치 비교를 위해 기준 되는 신호와 RMSE를 구한 것이다. 표에서 심층신경망을 이용해 구한 추정치의 RMSE가 가장 작음을 확인할 수 있다.

표 2. 잔향 환경 음원에 대한 사운드 보간 기법의 RMSE 결과

방법	RMSE	
	파형	스펙트럼 크기
산술평균	0.1608	0.1241
기하평균	0.1639	0.1264
DNN	0.0946	0.0762

4. 결론

UCC를 이용한 시청자 이동형 VR 음향구현에 있어서 두 지점 사이의 음향 신호를 생성할 때 심층신경망을 이용한 사운드 보간법에 대한 연구를 수행하였다. 우선, 두 지점에서 받은 음향 신호의 단구간 스펙트럼으로부터 통계적 방법인 산술평균과 기하평균을 이용해 사운드 보간을 진행해보았으나 실제 비선형 음향특성을 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.

감사의 글

이 논문은 2018년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임. [2016-0-00144, 시청자 이동형 자유시점 360VR 실감미디어 제공을 위한 시스템 설계 및 기반 기술 연구]

참고문헌

[1] V. Nair and G. E. Hinton, “Rectified linear units improve restricted Boltzmann machines,” in Proc. 27th Int. Conf. Machine Learning, pp. 807-814, 2010.

[2] Vu Pham, Théodore Bluche, Christopher Kermorvant, and Jérôme Louradour, “Dropout improves recurrent neural networksfor handwriting recognition,” Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference, pp. 285 - 290, IEEE, 2014.

[3] D. P. Kingma and J. L. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[4] C Veaux, J Yamagishi, K MacDonald. “SUPERSEDED-CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit”, 2016

[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. Y. C. Wen, N. D. Gaubitch, E. A .P. Habets, T. Myatt, and P. A. Naylor, “Evaluation of speech dereverberation algorithms using the MARDY database,” in Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC), Paris, France, 2006.