물질의 소리 특성 추론을 위한 컨볼루션 신경망

이도헌[1], 안인규[1], 임우빈[1], 윤성의[1]
[1] 한국과학기술원 전산학부

Acoustic Material Estimation with Convolutional Neural Network

Lee Doheon[1], An Inkyu[1], Im Woobin[1], Yoon Sung-Eui[1]
[1]KAIST School of Computing
e-mail: doheonlee@kaist.ac.kr, inkyu.an@kaist.ac.kr, iwbn@kaist.ac.kr, sungeui@kaist.edu

요 약

로봇이 활용할 수 있는 정보로서 청각 정보의 위상이 커지고 있다. 시각 정보와 더불어 장애물 탐지, 음성인식 등 많은 분야에서 로봇의 청각 정보가 활용되고 있다. 특히 물질의 소리 특성을 추정하는 것은 음성 인식, 공간 인식 등 다양한 분야의 발전을 위한 필수 요소이다. 실제 환경의 소리 특성을 추정하기 위하여 많은 연구가 진행되고 있지만, 주변 환경을 구성하는 물질의 다양성 때문에 어려움을 겪고 있다. 본 논문에서는 로봇이 공간적 특성을 인지하고, 물질의 차이에 따른 소리의 공간적 변화를 분류할 수 있도록 컨볼루션 신경망을 활용한 이진 분류기를 제안하였다. 이진 분류 실험 결과 83.9%의 정확도로 공간적 특성을 분류해 낼 수 있었다.

1. 서론

로봇의 주변 환경에 대한 인식은 로봇공학의 중요한 분야로서 연구됐다. 주변 환경 지도를 제작하며 로봇의 위치를 추정하는 Simultaneous Localization And Mapping (SLAM)은 현재까지도 많은 연구가 진행되고 있는 분야이다. 최근 스마트폰과 스마트 스피커의 보급으로 인하여 사용자 음성으로 로봇을 제어하는 음성 로봇 서비스는 중요한 분야로 성장하였다. 이에 따라 시각 정보뿐만 아니라, 청각 정보의 중요성 또한 증대되었고, 로봇이 소리를 인지하고 분석할 수 있는 능력을 향상시키기 위한 많은 연구가 진행되고 있다[1, 2].

소리는 주변 환경과 상호작용을 하며 전파된다. 물질의 종류에 따라서 소리를 반사하는 정도가 다르므로 반사율, 산란계수와 같은 물질의 음향 특성을 추정할 수 있어야 공간의 소리 전파 특성을 파악할 수 있다. 하지만 이러한 음향 특성은 물질의 종류와 소리의 주파수 대역에 따라 다르므로 실제 환경을 구성하는 모든 물질의 음향 특성을 정확히 추정하는 것은 매우 어려운 것으로 알려져 있다.

물질의 청각적 특성은 소리가 물질과 상호작용을 하는 전후의 변화로 정의할 수 있다. 다양한 물질에 대하여 음원에서 발생한 소리와 로봇에 측정된 소리의 연관성을 학습할 수 있다면, 일반적인 물질의 청각적 특성도 추정할 수 있다. 학습에는 인공신경망과 같은 기계학습 모델을 사용할 수 있다. 기계학습 모델을 활용할 때는 물질과 상호작용 전후 소리를 입력으로 두 입력 사이의 변화를 학습함으로써 물질의 성질을 추정할 수 있다.

본 논문에서는 공간을 구성하는 물질에 따른 소리의 변화를 로봇을 통해 구별할 수 있는지 검증하기 위하여, 컨볼루션 신경망을 이용한 소리의 이진 분류기를 제안한다. 실험 결과, 제안하는 이진 분류기는 높은 성능으로 두 가지 물질의 소리 특성을 분류해 내어 제안 모델의 타당성을 입증하였다.

2. 본론

2.1 스펙트로그램

소리를 푸리에 변환 (Fourier transform)을 통해 여러 개의 주파수의 파형으로 분리함으로써 고유한 특성을 나타낼 수 있다. 소리의 특성을 가시화하기 위하여 푸리에 변환된 신호를 시간을 가로축으로 하고 주파수를 세로축으로 하는 도면에 진폭을 밝기로 표현한 것이 소리의 스펙트로그램이다.

스펙트로그램은 각 주파수 대역에 대한 진폭 정보를 나타내고 있으므로 주파수 대역에 따라 다른 음향 특성을 추정하기에 용이하다. 또한, 회색스케일 이미지 형태의 자료이기 때문에 상대적으로 많은 연구가 진행된 컴퓨터 비전 분야의 연구성과들을 이용하여 정확한 소리 특성을 추출할 수 있다. 따라서 본 논문에서는 측정된 소리를 스펙트로그램으로 변환하여 활용하였다[그림 1].

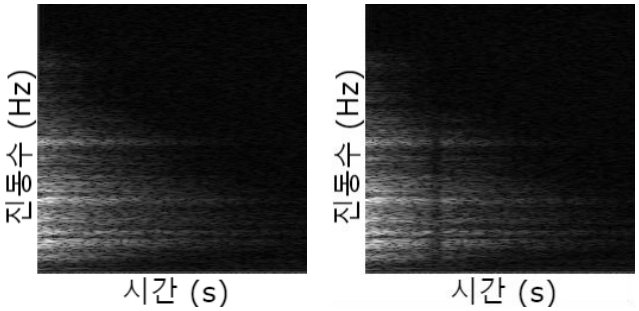
2.2 컨볼루션 신경망

이미지 분류, 물체 인식 등의 분야는 컨볼루션 인공 신경망을 활용함으로써 많은 발전을 이루었다. 시각 정보뿐만 아니라 청각 정보를 분석하는 연구에도 딥 러닝을 활용하는 연구들이 활발히 이루어지고 있다[3]. 본 논문에서는 청각 정보를 분석하는 방법으로서 물질에 따른 소리를 분류하기 위한 컨볼루션 이진 분류기를 구현하였다. 이진 분류기는 푸리에 변환을 통해 변환된 소리의 스펙트로그램을 입력값으로 받아 두 가지 물질에 대한 확률을 출력값으로 나타낸다.

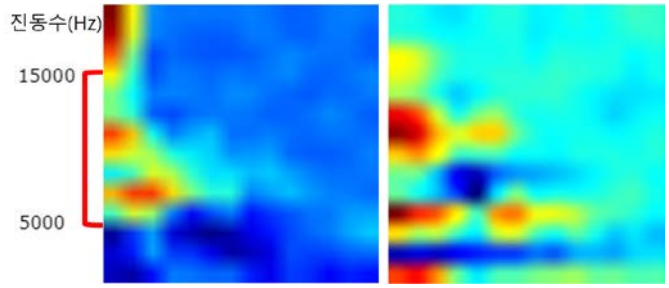
2.3 소리의 이진 분류 실험

실험 환경: 공간을 구성하는 물질에 따른 소리의 변화를 관측하기 위하여, [그림 3]과 같은 실험 환경을 구성하였다. 실험에서 물질 특성의 반영 정도에 변화를 주기 위하여 관측자와 물질의 거리(21.5cm, 30.0cm, 50.0cm)를 다르게 하며 실험을 진행하였다.

또한, 물질의 특성에 차이를 주고자 소리의 반사가 적은 물질 (딱딱한 면)과 소리의 반사가 큰 물질(부드러운 면)로 변인을 통제하였다. 음원, 관측자의 위치 등 다른 조건을 동일하게 한 후 물질의 종류를 바꿔가며 각각 관측자의 위치에서 측정하였다. 실험에 사용된 음원은 실험 공간이 아닌 곳에서 녹음한 박수 소리를 활용하였다.



[그림 1] 좌: 딱딱한 표면이 있는 환경에서 측정된 소리의 스팩트로그램. 우: 부드러운 표면이 있는 환경에서 측정된 소리의 스팩트로그램. 본 그림에 나타낸 이미지는 컨볼루션 신경망의 학습 데이터로서 사용된다.



[그림 2] [그림 1]의 스팩트로그램을 Class Activation Map (CAM)을 통해 확인한 신경망의 추론 영역. 신경망이 이미지의 어떤 부분을 근거로 분류하였는지 보여준다. 붉은색의 영역은 신경망의 높은 활성을 나타낸다. 좌: 딱딱한 표면의 결과. 우: 부드러운 표면의 결과.

이진 분류기 학습: 환경에서 측정된 소리의 처음부터 0.5초가 지난 후의 구간을 스팩트로그램으로 변환하였다. 변환된 2차원 이미지는 어떤 환경에서 측정되었는지에 대한 이진 라벨과 함께 컨볼루션 인공신경망의 학습 데이터로 활용되었다. 데이터 세트는 총 224개의 데이터로 구성되어 있으며 물질별 112개의 데이터가 존재한다. 이진 분류기는 전체 데이터를 활용한 8-폴드 교차검증을 통해 학습 및 평가되었다 [표 1].

유형	훈련	테스트	정확도
8-폴드 교차검증	196	28	0.8393±0.019

[표 1] 이진 분류기의 평가 유형 및 정확도

결과 분석: 소리의 이진 분류 실험 결과 83.9%의 정확도로 물질을 분류하였다. 이는 인공신경망이 높은 성능으로 물질에 따른 소리의 차이를 인지할 수 있음을 의미한다. 신경망이 어떻게 물질의 차이를 구분해 내는지 확인하기 위하여, Class Activation Map (CAM)을 통해 신경망의 활성을 나타내었다 [그림 2]. CAM을 통하여 신경망을 분석한 결과 초기 반사가 주요한 0.1초 이전의 소리에 높은 활성을 보였으며 주로 5,000Hz ~ 15,000Hz 주파수 대역의 소리에 높은 활성을 나타내었다. 이를 통해 신경망이 물질의 특성을 추론하기 위해 소리의 초기 반사 및 주파수 대역 별 물질의 소리 특성을 고려하고 있음을 확인할 수 있었다.

3. 결론

본 논문에서는 실험을 통해 공간의 물질 변화에 따른 소리의 변화를 로봇이 인지해낼 수 있음을 확인하였다. 이는 물질의 소리 특성을 분석하기 위하여 컨볼루션 신경망을 활용하는 것의 타당성을 입증해 주었다.

추후 음원의 종류와 상관없이 특성을 학습할 수 있도록 신경망을 구현할 예정이며, 물질의 반사율 및 산란 계수를 추론하는 등 물질의 청각적 특성을 추정하는 연구를 계획하고 있다.

[그림 3] 소리의 이진 분류를 위한 실험 환경. 좌: 딱딱한 표면이 있는 환경. 우: 부드러운 표면이 있는 환경.

후기

이 논문은 SW컴퓨팅산업원천기술개발사업 (SW스타랩, IITP-2015-0-00199)과 한국연구재단-차세대정보·컴퓨팅기술개발사업의 지원으로 수행되었음 (No. NRF2017M3C4A7066317).

참고문헌

[1] An, I., Son, M., Manocha, D., & Yoon, S. E. (2018, May). Reflection-Aware Sound Source Localization. In 2018 IEEE International Conference on Robotics and Automation (ICRA) (pp. 66-73). IEEE.

[2] Schissler, Carl, Christian Loftin, and Dinesh Manocha. "Acoustic classification and optimization for multi-modal rendering of real-world scenes." IEEE transactions on visualization and computer graphics 24.3 (2018): 1246-1259.

[3] Arandjelovic, R., & Zisserman, A. (2017, October). Look, listen and learn. In 2017 IEEE International Conference on Computer Vision (ICCV) (pp. 609-617). IEEE.