비균등 선형 마이크로폰 어레이를 활용한 합성곱 신경망 기반의 음원분리

문정민, 박인영, 김홍국

광주과학기술원 전기전자컴퓨터공학부

{jungmin7757, pinyoung, hongkook}@gist.ac.kr

Convolutional Neural Network Based Source Separation Using a Non-uniform Linear Microphone Array

Jung Min Moon, In Young Park , Hong Kook Kim

School of Electrical Engineering and Computer Science
Gwangju Institute of Science and Technology (GIST)

요약

본 논문에서는 비균등 선형 마이크로폰 어레이를 활용한 convolutional neural network (CNN) 기반의 음원분리 방법을 제안한다. 우선, 주어진 어레이 배치에 따라 채널간의 시간차를 분석하고, 분석된 시간차에 따라 주파수별로 방사각과 넓이에 따라 입력 오디오 신호의 spectral magnitude를 예측한다. 그러고 나서, CNN 분류기로부터 최적의 방사각과 넓이를 선별하고 이를 통해 음원을 분리한다.

1. 서론

음원분리 기술이란 여러 개의 객체 오디오 음원이 혼합된 오디오 신호에서 특정한 객체 오디오 신호만을 분리하는 기술을 말한다. 다양한 음원분리 기술 중에서, 최소 분산 무손실 응답(Minimum Variance Distortionless Response, MVDR) 빔형성기에서는 특정 방향에 빔을 형성함으로써 음원을 분리한다[1]. 하지만, 이러한 MVDR 기반 음원분리 기술은 스테레오 채널을 기반으로 수행되고 있으며 음원분리 기술로는 다소 부족한 면이 있다. 이러한 단점을 해결하기 위하여 스테레오 채널 기반의 음원분리 기술에서 다채널 기반의 비균등 선형 마이크로폰 어레이 기반의 음원분리가 시도되어 왔다[2]. 하지만 이 기술 또한 최적의 방사각(azimuth)과 넓이(width) 파라미터를 설정하는데 있어서 많은 어려움이 있다.

본 논문에서는 비균등 선형 마이크로폰 어레이 환경에서 convolutional neural network (CNN) 기반의 음원분리 방법을 제안한다. 제안된 방법에서는 비균등 선형 마이크로폰 어레이에 맞게 채널간의 시간차를 분석하고, 분석된 시간차에 상응하는 azimuth-frequency (AF) plane을 생성한다. 그러고 나서, 생성된 AF plane으로부터 CNN 기반의 분류기를 활용하여 추출하고자 하는 객체 오디오와 가장 일치하는 azimuth 및 width를 조절하여 음원분리를 수행한다.

2. 제안된 CNN 기반의 음원분리 방법

본 논문에서 제시한 CNN 기반의 음원분리의 전체 구성도는 <그림 1>과 같다. 영상 혹은 다른 미디어 정보로부터 들어온 객체 오디오의 위치를 이용하여 그 위치 정보 근처에 azimuth와 다양한 width로 하여 음원분리를 실행한다. 그 다음으로, 주어진 객체 오디오 정보를 통해 만들어진 CNN 기반의 분류기를 통하여 객체 오디오 정보와 분리된 음원들 간의 유사성을 측정한다. 이때 가장 큰 유사성을 갖는 음원이 최적의 azimuth과 width를 가지고 분리된 음원이라고 할 수 있다.

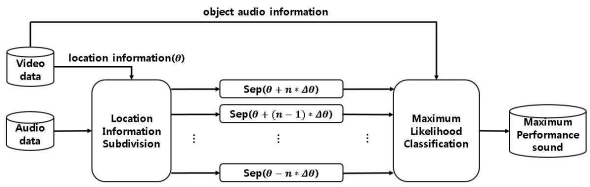

그림 1. 제안된 CNN 기반 음원분리 방법의 전체 구성도.

M개의 채널로 형성된 비균등 선형 마이크로폰 어레이를 활용하여 입력된 신호는 아래와 같이 표현될 수 있다.

$$\begin{bmatrix} x_1(n) \\ \vdots \\ x_M(n) \end{bmatrix} = \begin{bmatrix} a_1 s(n-\tau_1) \\ \vdots \\ a_M s(n-\tau_M) \end{bmatrix} + \begin{bmatrix} v_1(n) \\ \vdots \\ v_M(n) \end{bmatrix} \quad (1)$$

여기서, $s(n)$은 입력 신호로부터 분리하고자 하는 타겟신호를 의미하고, $x_i(n)$과 $v_i(n)$은 i번째 마이크로폰으로 들어오는 입력신호와 노이즈를 각각 의미한다. 또한, a_i와 τ_i는 타겟신호가 i번째 마이크로폰으로 입력될 때 감쇄와 지연 시간을 각각 나타낸다. 수식 (1)에 short time fourier transform (STFT)을 적용하여 주파수 영역으로 변환하면 아래 수식과 같다.

$$\boldsymbol{X} = \boldsymbol{d} S(k) + \boldsymbol{V} \quad (2)$$

여기서, $S(k)$는 $s(n)$의 k번째 주파수 성분이며, $\boldsymbol{X}^T$와 $\boldsymbol{V}^T$는 $[X_1(k), \cdots, X_M(k)]$와 $[V_1(k), \cdots, V_M(k)]$을 각각 나타낸다. 또한, $\boldsymbol{d}$는 $s(n)$을 마이크로폰 어레이로 입력받을 때, 방위각에 따라 나타나게 되는 감쇄와 지연시간을 각각 표현하는 벡터이다. 이때, 마이크로폰 어레이에 입력되는 신호가 far-field라고 가정한다면 $\boldsymbol{d}$는

$$\boldsymbol{d}^T = \left[W_N^{k\tau_1} \ \cdots \ W_N^{k\tau_M} \right] \quad (3)$$

위 수식과 같다. 여기서, $W_N^{k\tau_i} = \exp(-j2\pi k\tau_i / N)$이고 τ_i는 마이크로폰의 위치 및 타겟신호의 방향, θ에 따라서 결정되어지기 때문에 τ_i는 $\tau_i(\theta)$로 표현할 수 있다. 이 경우, $W_N^{k\tau_i(\theta)}$는 마이크로폰 위치와 타겟신호에 대하여 시간차를 보정하는 수치로 볼 수 있다. 이를 이용하는 AF plane은 다음과 같이 표현할 수 있다.

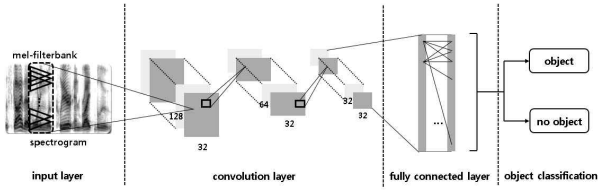

그림 2. CNN 기반 분류기 전체 구조도.

$$AF(k,\theta) = \left| W_N^{k\tau_1(\theta)} X_1(k) + \cdots + W_N^{k\tau_M(\theta)} X_M(k) \right| \quad (4)$$

여기서, 계산을 위해 θ를 sampling해야 한다. 본 논문에서는 AF plane의 resolution과 계산량을 고려하여 θ를 1°단위로 계산한다.

수식 (4)에서, 실제 타겟신호의 방향이 θ에 근접할수록 $AF(k,\theta)$가 커지게 된다. 즉, $AF(k,\theta)$가 최대가 되는 θ에서 타겟신호가 있다고 추정할 수 있다. 이에 근거하여, 주파수별로 최대값이 나타난 θ를 제외한 나머지 θ에 대한 $AF(k,\theta)$를 0으로 설정함으로써 θ에 따라 음원을 분리할 수 있다. 즉,

$$\overline{AF}(k,\theta) = \begin{cases} AF^{\max}(k), & \text{if } AF(k,\theta) = AF^{\max}(k) \\ 0, & otherwise \end{cases} \quad (5)$$

여기서, $AF^{\max}(k) = \max_\theta AF(k,\theta)$이다. 수식 (5)을 통하여 타겟신호의 θ에 따라 음원이 분리되고, azimuth, d_a 및 width, B의 설정에 따라 원하는 음원만을 다음식과 같이 추출한다.

$$|Y(k)| = \sum_{\theta = d_a - (B/2)}^{d_a + (B/2)} \overline{AF}(k,\theta) \quad (6)$$

여기서, 어떤 방위각에 해당하는 신호를 분리할 것인지는 d_a를 통해서 결정되며, 얼마만큼의 방위각 넓이를 설정할 것인지는 B를 통해서 결정된다. 즉, 수식 (6)으로부터 획득한 magnitude 성분과 원음의 phase 성분을 가지고 최종적으로 객체 오디오 신호를 분리한다. 이때 최적의 d_a 및 B를 설정하는 문제가 발생하며, 이를 위해 CNN 기반의 분류기를 설계하였다.

CNN 기반의 분류기는 <그림 2>와 같이 나타낼 수 있다[3]. 그림에서 보는 바와 같이, 본 논문에서 CNN은 입력층 1층, 합성곱 2층, fully connected 1층 그리고 object classification으로 구성되어 있다. 먼저, 입력층에서는 특징 추출을 위하여 STFT를 적용한다. 여기서, FFT size는 256로, overlap length은 128로 각각 설정하였다. 이 주파수성분에 mel-feature를 얻기 위해서 128 밴드의 mel-filterbank를 적용한다. 이와 같은 방식으로부터 얻어진 이미지의 크기는 총 이미지의 수를 N이라고 한다면 $N \times 128 \times 32 \times 1$가 된다. 본 논문에서는 총 이미지의 수, N은 9,032이다.

이 이미지들을 축소하기 위해서 합성곱 두 개의 층에서 max-pooling(2×2)을 진행한다. 결과적으로 총 이미지의 크기는 $(N/4) \times 32 \times 32 \times 1$가 된다. 이 때, 활성화 함수는 ReLu를 사용하며, 과학습 현상이 발생할 수 있으므로 각 층마다 드랍아웃을 적용한다. 그리고 나서, 이미지를 모두 붙여서 fully connected layer를 만든다. 그리고, fully connected layer의 데이터를 토대로 object를 분류할 수 있다.

이 분류기를 활용하기 위하여 타겟신호의 방위각 주변에 있는 azimuth들과 여러가지 width들을 설정하여 여러 개의 객체화된 음원을 추출한다. 그 다음으로, 학습된 분류기로부터 타겟신호와의 유사성을 측정한다. 이때 가장 유사성이 높게 나타난 음원에 상응하는 azimuth와 width를 최적으로 판단하여 음원을 분리한다.

본 논문에서 제안된 방법은 다음과 같은 장점이 있다. 첫째로, 혼합된 가중치와 관계없이 최적의 azimuth와 width를 찾을 수 있으며, 둘째로는 움직이는 오디오 객체에 대해서도 최적의 azimuth와 width를 쉽게 찾을 수 있다.

표 1. 각 방법에 따른 음원별 분리정확도(%) 비교.

방법 / 음원	MVDR	Non-uniform Linear Microphone Array Method	Proposed Method
Violin	73.6	97.0	97.0
Clarinet	77.9	76.4	78.5
Cello	78.1	82.7	87.1
Avg.	76.5	85.3	87.3

3. 성능평가

제안된 방법의 성능을 평가하기 위해서 분리정확도를 측정하였다. 분리정확도는 음원 분리된 신호가 실제 reference 음원 중에서 어떤 음원과 가장 유사한지 프레임별로 판단하여 이에 대해 통계적으로 수치화한 것이다[4]. 본 실험에서는 총 8채널의 비균등 선형 마이크로폰 어레이를 활용하였으며, 음원은 총 3가지 악기(violin, clarinet, cello)로 구성된 오케스트라 연주곡으로 연주자의 위치를 -45°, 0°, 45°로 각각 배치하여 녹음하였다. 이 때 음원의 길이는 120초로 구성하였으며, 연주자들의 연주로 인한 움직임으로 소리의 방사각은 약간씩 변화하였다. 여기서, 객체 오디오의 정보는 영상 혹은 다른 미디어 정보로부터 알 수 있다고 가정하였다.

<표 1>은 분리정확도의 측정된 결과를 보여준다. 제안된 방법의 분리정확도가 MVDR 빔형성기에 비하여 10.8%, 비균등 선형 마이크로폰 어레이 기반의 음원분리에 비하여 2.0% 높게 나타나는 것을 확인하였다.

4. 결론

본 논문에서는 비균등 선형 마이크로폰 어레이 환경에서 CNN 분류기를 활용한 음원분리 기술을 제안하였다. 제안된 음원분리 방법은 8채널 마이크로폰을 활용하여 채널간의 시간차 분석을 통해 AF plane을 생성하였다. 생성된 AF plane으로부터 CNN 기반의 분류기를 활용하여 방위각 및 넓이를 조절하여 음원분리를 수행하였다. 제안된 방법은 MVDR 빔형성기와 비균등 선형 마이크로폰 어레이 기반의 음원분리 방법에 비해 각각 10.8%와 2.0%의 음원분리 정확도 향상을 보였다.

감사의 글

본 연구는 과기정통부 및 정보통신기술진흥센터의 정보통신·방송 연구개발 사업[R0126151034002003, 채널/객체 융합형 하이브리드 오디오 콘텐츠 제작 및 재생기술 개발]과 한국연구재단의 지원을 받아 수행된 연구임(No. 2015R1A2A1A05001687).

참고 문헌

[1] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 10, pp. 1365-1375, Oct. 1987.

[2] C. J. Chun and H. K. Kim, "Non-uniform linear microphone array based source separation for conversion from channel-based to object-based audio content," *Journal of Broadcast Engineering*, vol. 21, no. 2, pp. 169-179, Mar. 2016.

[3] T. Park and T. Lee, "Musical instrument sound classification with deep convolutional neural network using feature fusion approach," *arXiv* preprint arXiv, 2015.

[4] A. L. Casanovas, G. Monaci, P. Vandergheynst, and R. Gribonval, "Blind audiovisual source separation based on sparse redundant representations," *IEEE Trans. Multimedia*, vol. 12, no. 5, pp. 358-371, Aug. 2010.