# 방송용 오디오 콘텐츠 제작을 위한 비균등 선형 마이크로폰 어레이 기반의 음원분리 방법

전찬준, 김홍국
광주과학기술원
{cjchun, hongkook}@gist.ac.kr

# Non-uniform Linear Microphone Array Based Source Separation for Broadcasting Audio Content Production

Chan Jun Chun, Hong Kook Kim
Gwangju Institute of Science and Technology (GIST)

## 요약

현재 UHDTV (Ultra-High-Definition TV) 시대에 사용될 멀티미디어 부호화로 MPEG-H를 표준화로 진행하고 있다. 향후 방송용 오디오 콘텐츠는 채널 오디오 콘텐츠에서 진화하여 객체 오디오 콘텐츠까지도 필요하게 된다. 이에 따라, 본 논문에서는 고품질의 방송용 오디오 콘텐츠를 제작하기 위한 비균등 선형 마이크로폰 어레이 기반의 음원분리 방법을 제안한다. 제안된 방법은 주어진 어레이 배치에 따라 채널간의 시간차를 분석하고, 이에 따른 객체 오디오 생성을 위한 음원분리 기술을 적용한다. 제안된 기법의 성능을 검증하기 위하여 음원분리도를 측정하였고, MVDR (Minimum Variance Distortionless Response) 빔형성기와 성능을 비교하였다. 비교 결과, 제안된 기법이 MVDR 빔형성기에 비하여 12.8% 높은 음원분리도 수치를 나타낸 것을 확인하였다.

## 1. 서론

최근 실감 비디오 기술과 더불어 오디오 기술에 관한 연구가 활발히 진행되고 있으며, 특히 UHDTV (Ultra-High-Definition TV) 시대에 사용된 멀티미디어 부호화로 MPEG-H가 표준화로 진행되고 있다 [1]. 특히, MPEG-H의 3D audio는 NHK 22.2채널 방송과 같은 실감 오디오 콘텐츠에 더불어 객체 오디오 콘텐츠까지도 지원하고, 이러한 콘텐츠를 위한 오디오 부호화 및 복호화 기술과 다양한 출력채널 환경에 적응할 수 있는 렌더링 (rendering) 기술을 표준화 대상으로 규정하고 있다 [1]. 향후 방송용 오디오 콘텐츠는 채널 오디오 콘텐츠에서 진화하여 객체 오디오 콘텐츠까지도 지원이 필요할 전망이며, 이에 따라 음원분리 기술을 통하여 채널 오디오 콘텐츠를 객체 오디오 콘텐츠로 변환하는 기술이 요구된다.

최근들어 음원분리 기술에 대한 여러 알고리즘들이 활발히 연구되어 오고 있다 [2][3]. 그 중에서도 독립 성분 분석 (independent component analysis, ICA) 알고리즘은 오디오 신호들간에 상호독립적이며, non-Gaussian이라는 가정을 통하여 음원을 분리한다 [2]. 반면, 계산 청각장면 분석(computational auditory scene analysis, CASA) 알고리즘에서는 인체 청각 시스템의 메커니즘을 기반으로 음원을 분리한다 [3]. 하지만, 이러한 음원분리 기술들은 스테레오 채널을 기반으로 수행되고 있으며, 방송용 오디오 콘텐츠 제작을 위하여 다채널의 오디오 콘텐츠를 고품질 객체 오디오 콘텐츠로 변환하기 위한 음원분리 기술로는 다소 부족한 면이 있다.

따라서 본 논문에서는 고품질의 방송용 오디오 콘텐츠를 제작하기 위한 비균등 선형 마이크로폰 어레이 기반의 음원분리 방법을 제안한다. 제안된 방법에서는 비균등 선형 마이크로폰 어레이에 맞게 채널간의 시간차를 분석하고, 분석된 시간차에 상응하는 azimuth-frequency (AF) plane을 생성한다. 이후, 생성된 AF plane으로부터 azimuth 및 width를 조절하여 음원분리를 수행한다.

## 2. 제안된 비균등 선형 마이크로폰 어레이 기반의 음원분리 방법

$M$개의 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, $x_i(n)$과 $v_i(n)$은 $i$번째 마이크로폰으로 수음되는 입력신호와 노이즈를 각각 의미하며, $s(n)$은 입력 신호로부터 분리하고자 하는 타겟 신호이다. 또한, $a_i$과 $\tau_i$는 타겟 신호가 $i$번째 마이크로폰으로 입력될 때 감쇄와 지연 시간을 각각 나타낸다. 수식 (1)을 $N$-point fast Fourier transform (FFT)를 통하여 주파수 영역으로 변환하면 아래의 수식과 같다.

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

여기서, $\boldsymbol{X}^{\boldsymbol{T}}$와 $\boldsymbol{V}^{\boldsymbol{T}}$는 $[X_1(k) \cdots X_M(k)]$와 $[V_1(k) \cdots V_M(k)]$이며, $S(k)$는 $s(n)$의 $k$번째 주파수 성분을 나타낸다. 또한, $\boldsymbol{d}$는 $s(n)$의 마이크로폰 어레이로 입력받을 때, 방위각에 따라 나타나게 되는 감쇄와 지연시간을 표현하는 벡터이다. 즉,

$$\boldsymbol{d}^{\boldsymbol{T}} = \left[ a_1 \exp(-j\frac{2\pi k\tau_1}{N}) \ \cdots \ a_M \exp(-j\frac{2\pi k\tau_M}{N}) \right] \quad (3)$$

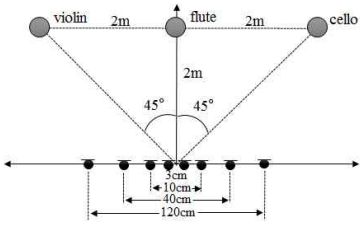


그림 1. 음원분리도 측정을 위한 비균등 선형 마이크로폰 어레이 및 음원 배치 환경

마이크로폰 어레이에 입력되는 신호가 far-field 모델이라고 가정할 때에는 감쇄 인자는 모두 동일하다는 가정하에, 수식 (3)은 아래와 같이 간략화된다 [4].

$$\boldsymbol{d}^{T}=\left[\, W_N^{k\tau_1} \;\cdots\; W_N^{k\tau_M} \right] \tag{4}$$

여기서, $W_N^{k\tau_i}=\exp(-j2\pi k\tau_i/N)$이다. 수식 (4)에서 $\tau_i$는 마이크로폰의 위치 및 타겟 신호의 방향 $\theta$에 따라서 결정되어지기 때문에 $\tau_i$는 $\tau_i(\theta)$로 표현할 수 있다. 이 경우, $W_N^{k\tau_i(\theta)}$는 마이크로폰 위치와 타겟 신호에 대하여 시간차를 보정하는 수치로 볼 수 있다. 이를 이용하여 AF plane을 아래의 수식처럼 정의할 수 있다.

$$AF(k,\theta)=\left| W_N^{k\tau_1(\theta)}X_1(k)+\cdots+W_N^{k\tau_M(\theta)}X_M(k)\right| \tag{5}$$

여기서, 계산을 위해 $\theta$를 sampling해야 하는 데, 본 논문에서는 AF plane의 resolution과 계산량을 고려하여 1° 단위로 $\theta$을 계산하였다.

수식 (5)에서 보는 바와 같이, 실제 타겟 신호의 방향이 $\theta$에 근접할수록 $AF(k,\theta)$ 값이 커지게 된다. 즉, $AF(k,\theta)$가 최대가 되는 $\theta$에서 타겟 신호가 있다고 추정할 수 있다. 이에 근거하여, 주파수별로 최대가 되는 $\theta$를 제외한 나머지 $\theta$에 대한 $AF(k,\theta)$를 0으로 설정함으로써 $\theta$에 따라서 음원들을 분리할 수 있다.

$$\overline{AF}(k,\theta)=\begin{cases} AF^{\max}(k), & \text{if } AF(k,\theta)=AF^{\max}(k) \\ 0, & otherwise \end{cases} \tag{6}$$

여기서, $AF^{\max}(k)=\max_{\theta}AF(k,\theta)$이다. 수식 (6)을 통하여 타겟 신호의 각각의 방향별로 $\theta$에 따라서 음원들이 분리되고, azimuth, $d_a$ 및 width, $B$의 설정에 따라서 원하는 음원만을 추출할 수 있다. 즉,

$$|Y(k)|=\sum_{\theta=d_a-(B/2)}^{d_a+(B/2)}\overline{AF}(k,\theta) \tag{7}$$

수식 (7)에서와 같이, 어떤 방위각에 해당하는 신호를 분리할 것인지는 $d_a$를 통해서 결정되어 지며, 얼마만큼의 방위각 넓이로 분리할 것인지는 $B$를 통해서 결정된다. 마지막으로, 수식 (7)로 획득한 magnitude 성분과 원음의 phase 성분을 가지고 최종적으로 객체 오디오 신호를 분리한다.

## 3. 성능평가

제안된 방법의 성능을 평가하기 위해서 음원분리도를 측정하였다 [5]. 음원분리도는 음원 분리된 신호가 실제 reference 음원 중에서 어떤 음원과 가장 유사한지 프레임별로 판단하여 이에 대해 통계적으로 수치화한 것이다 [5]. 음원분리도 측정을 위하여 <그림 1>과 같은 환경을 설정하였다. 고품질의 음원분리를 위하여 총 8채널의 비균등 선형 마이크로폰을 활용하였으며, 이에 따라서 총 3개의 음원(바이올린, 플룻, 첼로)을 -45°, 0°, 45° 각각 배치하였다. 성능 비교를 위해, minimum variance distortionless response (MVDR) 빔형성기[4]를 적용한 후 그 음원분리도를 함께 측정하였다.

표 1. 음원 종류에 따른 제안된 방법의 음원분리도 및 MVDR과 비교

| 음원 종류 | MVDR (%) | 제안된 방법 (%) |
|---|---|---|
| Violin | 39.8 | 63.6 |
| Flute | 77.9 | 91.7 |
| Cello | 93.4 | 94.2 |
| 평균 | 70.4 | 83.2 |

<표 1>은 음원분리도의 측정 결과를 보여준다. 표의 결과를 통해, 제안된 음원분리 방법의 음원분리도가 MVDR 빔형성기에 비하여 12.8% 높게 나타나는 것을 알 수 있었다.

## 4. 결론

본 논문에서는 비균등 선형 마이크로폰 어레이 환경에서 고품질의 방송용 오디오 콘텐츠 제작을 위한 음원분리 기술을 제안하였다. 제안된 음원분리 방법은 8채널 마이크로폰을 활용하여 채널간의 시간차를 분석하고, 이에 기반하여 AF plane을 생성하였다. 생성된 AF plane으로부터 azimuth 및 width를 조절하여 magnitude를 추정하고 이를 통해 음원분리를 수행하였다. 제안된 방법의 성능을 검증하기 위하여 MVDR 빔형성기와의 음원분리도를 비교하였다. 그 결과, 제안된 방법이 MVDR 빔형성기에 비하여 12.8% 높은 음원분리도를 보였다.

## 감사의 글

본 연구는 2015년도 미래창조과학부 및 정보통신기술진흥센터의 정보통신·방송 연구개발 사업 [R01261510340002003, 채널/객체 융합형 하이브리드 오디오 콘텐츠 제작 및 재생기술 개발]과 한국연구재단의 지원을 받아 수행된 연구임 (No. 2015R1A2A1A05001687).

## 참고 문헌

[1] J. Herre, J. Hilpert, A. Kuntz, and J, Plogsties, “MPEG-H 3D audio—the new standard for coding of immersive spatial audio,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 770-779, Aug. 2015.

[2] A. Hyvärinen, “Survey on independent component analysis,” *Neural Computing Surveys*, vol. 2, pp. 94-128, 1999.

[3] D. F. Rosenthal and H. G. Okuno, *Computational Auditory Scene Analysis*, LEA Publishers, Mahwah, NJ, 1998.

[4] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, Springer, Berlin, Germany, 2008.

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