특집논문 (Special Paper)
방송공학회논문지 제21권 제2호, 2016년 3월 (JBE Vol. 21, No. 2, March 2016)
http://dx.doi.org/10.5909/JBE.2016.21.2.169
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

채널 기반에서 객체 기반의 오디오 콘텐츠로의 변환을 위한 비균등 선형 마이크로폰 어레이 기반의 음원분리 방법

전 찬 준[a)], 김 홍 국[a)‡]

Non-uniform Linear Microphone Array Based Source Separation for Conversion from Channel-based to Object-based Audio Content

Chan Jun Chun[a)] and Hong Kook Kim[a)‡]

요 약

오늘날 UHDTV (Ultra-High-Definition TV) 시대에 사용될 멀티미디어 부호화기로 MPEG-H에 대한 표준화가 진행되고 있다. 향후 방송용 오디오 콘텐츠는 채널 기반 오디오 콘텐츠에서 진화하여 객체 기반 오디오 콘텐츠까지도 포함하게 될 예정이다. 이에 따라, 채널 기반 오디오 콘텐츠의 객체 기반 오디오 콘텐츠로의 유기적인 변환이 필요한 실정이다. 본 논문에서는 이러한 유기적인 변환을 실현 가능하게 할 수 있는 비균등 선형 마이크로폰 어레이 기반의 음원분리 기법을 제안한다. 제안된 기법은 주어진 어레이 배치에 따라 채널간의 시간차를 분석하고, 분석된 시간차에 따라 주파수별로 특정 방위각에 위치한 입력 오디오 신호의 spectral magnitude를 예측한다. 이후, azimuth와 width 파라메타를 조정함으로써 객체 오디오 생성을 위한 음원을 분리한다. 제안된 음원분리 기법의 성능을 평가하기 위하여 객관적 음원분리 지표 및 분리정확도를 측정하였고, 최소 분산 무손실 응답 빔형성기와 독립 성분 분석 기법 등 기존 음원분리 기법과의 그 성능을 비교하였다. 비교 결과, 제안된 기법이 기존 음원분리 기법들에 비하여 우수한 음원분리 성능을 보이는 것을 알 수 있었다.

Abstract

Recently, MPEG-H has been standardizing for a multimedia coder in UHDTV (Ultra-High-Definition TV). Thus, the demand for not only channel-based audio contents but also object-based audio contents is more increasing, which results in developing a new technique of converting channel-based audio contents to object-based ones. In this paper, a non-uniform linear microphone array based source separation method is proposed for realizing such conversion. The proposed method first analyzes the arrival time differences of input audio sources to each of the microphones, and the spectral magnitudes of each sound source are estimated at the horizontal directions based on the analyzed time differences. In order to demonstrate the effectiveness of the proposed method, objective performance measures of the proposed method are compared with those of conventional methods such as an MVDR (Minimum Variance Distortionless Response) beamformer and an ICA (Independent Component Analysis) method. As a result, it is shown that the proposed separation method has better separation performance than the conventional separation methods.

Keyword : Sound source separation, channel-based audio content, object-based audio content, non-uniform linear microphone array, frequency-dependent source separation

Ⅰ. 서 론

최근 실감 비디오 기술과 더불어 오디오 기술에 관한 연구가 활발히 진행되고 있으며, 특히 UHDTV (Ultra-High-Definition TV) 시대에 사용될 멀티미디어 부호화기로 MPEG-H에 대한 표준화로 진행되고 있다[1]. 특히, MPEG-H의 3D audio는 NHK 22.2채널 방송과 같은 채널 오디오 콘텐츠와 더불어 객체 오디오 콘텐츠까지도 지원하기 위해 필요한 오디오 부호화 및 복호화 기술과 다양한 출력채널 환경에 적응할 수 있는 렌더링(rendering) 기술을 표준화 대상으로 규정하고 있다[1,2]. 이를 통해 객체 오디오 콘텐츠를 활용함으로써 사용자가 직접 원하는 위치에 객체 오디오를 정위(localization)할 수 있고, 볼륨 및 재생 여부까지 제어가 가능하다는 장점을 갖게 된다[1,2]. 따라서, 향후 방송용 오디오 콘텐츠는 채널 오디오 콘텐츠에서 진화하여 객체 오디오 콘텐츠로 전향될 전망이며, 객체 오디오 콘텐츠에 대한 수요를 충족시키기 위한 하나의 방법으로 기존의 채널 오디오 콘텐츠로부터 객체 오디오를 분리하는 방법이 그 대안으로 고려되고 있다.

음원분리 기술이란 여러 개의 객체 오디오가 혼합된 오디오 신호에서 특정 객체 오디오 신호만을 분리하거나 추출하는 기술을 의미한다[3]. 최근 들어 음원분리 기술에 대한 여러 알고리즘들이 활발히 연구되고 있다[4-7]. 그 중에서도 독립 성분 분석(independent component analysis, ICA) 알고리즘은 오디오 신호들간에 상호독립적이며 non-Gau-ssian이라는 가정을 통하여 음원을 분리한다[4]. 반면, 계산 청각장면 분석(computational auditory scene analysis, CASA) 알고리즘은 인체 청각 시스템의 메커니즘을 기반으로 음원을 분리한다[5]. 또한 최소 분산 무손실 응답(Minimum Variance Distortionless Response, MVDR) 빔형성기[6]에서는 특정 방향에 빔을 형성함으로써 음원분리를 수행하며, DUET (degenerate unmixing estimation technique) 알고리즘에서는 스테레오 오디오 신호에 혼합되어 있는 각각의 객체 오디오가 W-disjoint orthogonal하다고 가정하고 감쇄 및 지연에 대한 히스토그램을 생성하여 음원을 분리한다[7]. 하지만, 이러한 음원분리 기술들은 모노 혹은 스테레오 채널을 기반으로 수행되고 있으며, MVDR 빔형성기의 경우 노이즈 성분에 대한 suppression하는 성능을 기대할 수 있지만, 방송용 오디오 콘텐츠 제작을 위하여 다채널 오디오 콘텐츠를 고품질 객체 오디오 콘텐츠로 변환하기 위한 음원분리 기술로는 다소 부족한 면이 있다[8].

따라서, 본 논문에서는 고품질의 방송용 오디오 콘텐츠를 제작하기 위한 비균등 선형 마이크로폰 어레이 기반의 음원분리 기법을 제안한다. 제안된 기법에서는 비균등 선형 마이크로폰 어레이에 맞게 채널간의 시간차를 분석하고, 분석된 시간차에 상응하는 azimuth-frequency (AF) plane을 생성한다[9,10]. 이후, 생성된 AF plane으로부터 주파수별로 최대값이 되는 방위각에 대해서 입력 오디오 신호의 magnitude를 예측하게 된다. 그리고 나서, azimuth 및 width 파라메타를 조절함으로써 음원분리가 수행된다. 본 논문에서 제안된 음원분리 기법의 성능을 평가하기 위하여 공연 잔향이 존재하는 소극장 환경에서 연주자들이 악기를 연주하는 합주를 녹음 받고, 이를 제안된 음원분리 기법을 통하여 각각의 객체 오디오를 획득하였다. 음원분리 기술의 성능은 객관적 분리지표[11] 및 분리정확도 지표[12]로 측정하고, 기존 음원분리 기법인 ICA와 MVDR 빔형성기와의 그 성능을 비교한다.

본 논문의 구성은 다음과 같다. 서론에 이어, 2절에서는 기존의 음원분리 기법인 MVDR 빔형성기 및 ICA 기법에 대하여 설명하고, 3절에서는 비균등 선형 마이크로폰 어레이 기반의 음원분리 기법을 제안한다. 그리고, 4절에서는 제안된 음원분리 기법의 성능을 평가하기 위한 실험 환경과 성능 측정 방법을 기술하고 ICA와 MVDR 빔형성기와

a) 광주과학기술원 전기전자컴퓨터공학부(School of Electrical Engineering and Computer Science)
‡ Corresponding Author : 김홍국(Hong Kook Kim)
E-mail: hongkook@gist.ac.kr
Tel: +82-62-715-3121
ORCID: http://orcid.org/0000-0002-0105-6693
※ 본 연구는 2015년도 미래창조과학부 및 정보통신기술진흥센터의 정보통신·방송 연구개발 사업 [R0126151034000 2003, 채널/객체 융합형 하이브리드 오디오 콘텐츠 제작 및 재생기술 개발]과 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2015R1A2A1A05001687).
※ 이 논문의 연구결과 중 일부는 “2015년 한국방송공학회 추계학술대회”에서 발표한 바 있음.
· Manuscript received January 25, 2016; Revised March 22, 2016; Accepted March 22, 2016.

제안된 기법과의 성능을 비교한다. 마지막으로 5절에서는 본 논문의 결론을 맺는다.

Ⅱ. 기존의 음원분리 기법

1. MVDR 빔형성기

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, $x_i(n)$과 $v_i(n)$은 i번째 마이크로폰으로 수음되는 입력 오디오 신호와 노이즈 성분을 각각 의미하며, $s(n)$은 입력 신호로부터 분리하고자 하는 타겟 음원신호이다. 또한, a_i과 τ_i는 타겟 음원신호가 i번째 마이크로폰으로 입력될 때 감쇄와 지연 시간을 각각 나타낸다. 수식 (1)을 STFT (short-time Fourier transform)를 통하여 주파수 영역으로 변환하면 아래의 수식과 같다.

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

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

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

여기서, N은 STFT의 point를 가리킨다. 수식 (3)에서 τ_i는 객체 오디오의 방향, 소리의 속도, 그리고 표본화율에 따라서 결정 가능하다. 즉,

$$\boldsymbol{d}^{\boldsymbol{T}} = \left[a_1 \exp(-j\frac{2\pi k}{N}\frac{f_s}{c} l_1 \sin\theta) \ \cdots \ a_M \exp(-j\frac{2\pi k}{N}\frac{f_s}{c} l_M \sin\theta) \right] \quad (4)$$

여기서, f_s는 표본화율을 가리키며, c는 소리의 속도, θ는 객체 오디오의 방향, 그리고 l_i는 마이크로폰 어레이의 중심으로부터의 i번째 마이크로폰까지의 간격을 각각 의미한다.

마이크로폰 어레이의 간격이 입력되는 신호와의 거리에 비하여 충분히 가깝다고 가정하면 far-field 모델이라고 할 수 있고[13,14], 또한 감쇄 인자는 모두 동일하다는 가정 하에 수식 (4)는 아래의 수식처럼 간략화가 가능하다.

$$\boldsymbol{d}^{\boldsymbol{T}} = \left[W_N^{k\tau_1} \ \cdots \ W_N^{k\tau_M} \right] \quad (5)$$

여기서, $W_N^{k\tau_i} = \exp(-j2\pi k\tau_i/N)$이다. 따라서, 주파수 도메인에서 수행되는 일반적인 빔형성기는 아래의 수식처럼 입력 신호에 가중 벡터를 선형 조합하는 형태로 결정된다.

$$\hat{S}(k) = \boldsymbol{W}^{\boldsymbol{H}}\boldsymbol{X} = \left[W_1(k) \ \cdots \ W_M(k) \right] \begin{bmatrix} X_1(k) \\ \vdots \\ X_M(k) \end{bmatrix} \quad (6)$$

여기서, $\boldsymbol{W}$는 빔형성기의 가중 벡터를 나타내며, H는 Hermitian 연산자를 가리킨다.

MVDR 빔형성기는 원하는 방향에 대한 신호의 크기를 일정하게 유지하면서 나머지 방향에 대한 신호의 크기를 최소화하는 형태로 빔을 형성한다[6].

$$\min \boldsymbol{W}_M^{\boldsymbol{H}} \boldsymbol{R}_{\boldsymbol{XX}} \boldsymbol{W}_M \text{ subject to } \boldsymbol{W}_M^{\boldsymbol{H}} \boldsymbol{d} = 1 \quad (7)$$

여기서, $\boldsymbol{R}_{\boldsymbol{XX}}$는 입력된 오디오 신호의 자기상관행렬을 나타내며, $\boldsymbol{d}$는 수식 (5)에서와 같은 조향 벡터를 가리킨다. 수식 (7)의 조건을 만족하는 가중 벡터를 찾는 것이 MVDR 빔형성 기법의 핵심이며, 이는 Lagrange multiplier를 활용하여 아래의 수식과 같이 가중 벡터를 결정할 수 있다[13].

$$\boldsymbol{W}_M = \frac{\boldsymbol{R}_{\boldsymbol{XX}}^{-1}\boldsymbol{d}}{\boldsymbol{d}^{\boldsymbol{H}}\boldsymbol{R}_{\boldsymbol{XX}}^{-1}\boldsymbol{d}} \quad (8)$$

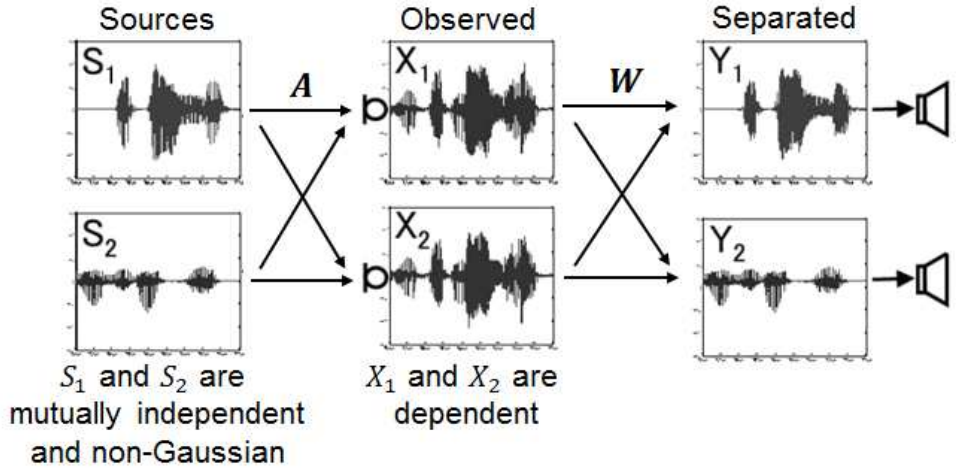


그림 1. 독립 성분 분석 기반의 음원분리 기법의 개요
Fig. 1. Overview of source separation based on independent component analysis (ICA)

상기의 수식에서와 같이 조향 벡터 $\boldsymbol{d}$를 원하는 방향으로 설정함으로써 비균등 선형 마이크로폰 어레이를 활용한 음원분리가 가능해진다.

2. ICA 기법

ICA 기법에서는 독립적인 객체 오디오 신호들을 <그림 1>에서 보인 바와 같이 정방 행렬 A를 통해서 혼합된 성분들이 입력 오디오 신호로 관측될 때 이는 아래의 수식으로 표현할 수 있다[4].

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s} \tag{9}$$

여기서, $\boldsymbol{x}$와 $\boldsymbol{s}$는 $[x_1(n) \cdots x_M(n)]^T$와 $[s_1(n) \cdots s_M(n)]^T$을 가리킨다. 입력된 오디오 신호인 $\boldsymbol{x}$만을 가지고 혼합 행렬인 A를 추정함으로써 음원분리가 이루어진다.

$$\hat{\boldsymbol{s}} = \boldsymbol{W}\boldsymbol{x} \tag{10}$$

ICA 기법에서는 혼합 행렬 $\boldsymbol{W}$를 예측하기 위하여 객체 오디오 신호들이 독립적인 특성을 가진다는 가정에 더해서 non-Gaussian 특성을 가진다고 가정한다[15]. 이에 따라, 아래 식과 같은 fourth-order cumulant인 kurtosis를 활용할 수 있다.

$$kurt(\boldsymbol{x}) = E\{\boldsymbol{x}^4\} - 3\left(E\{\boldsymbol{x}^2\}\right)^2 \tag{11}$$

여기서, kurtosis 값이 양수 값을 가질수록 super-Gaussian 분포를 갖는다는 것을 의미하며, 음수 값을 가질수록 sub-Gaussian 분포를 가지게 된다. Super-Gaussian이란 표준편차가 1인 Gaussian 분포보다 폭이 좁은 형태를 의미하며, sub-Gaussian이란 폭이 더 넓은 형태의 Gaussian 분포를 의미한다. Kurtosis를 최대 혹은 최소로 만드는 객체 오디오 신호가 가장 큰 non-Gaussian 특성을 가지게 되며, 이를 통해서 행렬 $\boldsymbol{W}$를 구할 수 있다[15].

Ⅲ. 제안된 비균등 선형 마이크로폰 어레이 기반의 음원분리 기법

1. 개요

본 논문에서 제안된 비균등 선형 마이크로폰 어레이 기반의 음원분리 기법의 구성도는 <그림 2>와 같다. 먼저, M개 채널로 구성된 마이크로폰 어레이로부터 오디오 신호를 입력 받는다. 입력된 오디오 신호는 windowing 및 STFT를 통하여 주파수 영역으로 변환된다. 본 논문에서는 50%

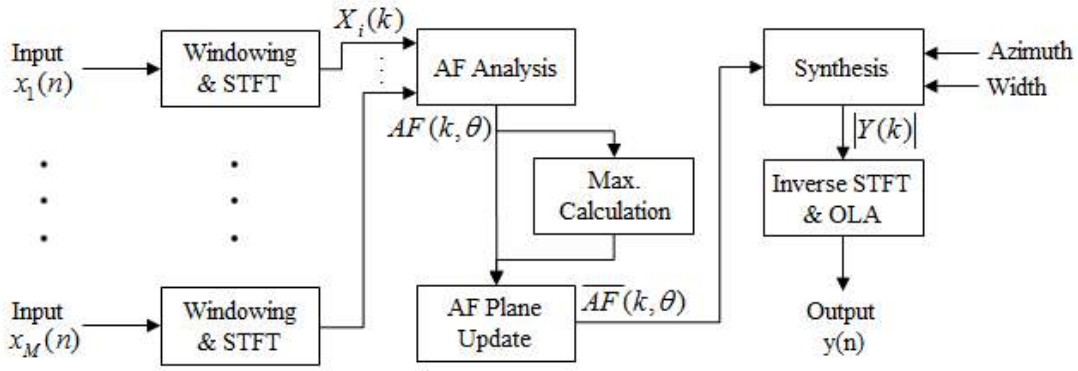


그림 2. 제안된 비균등 선형 마이크로폰 어레이 기반의 음원분리 기술의 구성도
Fig. 2. Procedure of the proposed non-uniform linear microphone array based source separation technique

overlap된 hanning window를 적용하여 4096-point의 STFT을 수행한다. 주파수 영역으로 변환된 오디오 신호는 방위각별로 시간차를 보정하고, 각 방위각과 주파수의 함수로 AF plane을 생성한다. 생성된 AF plane으로부터 주파수별로 최대값이 되는 방위각에 대해서 입력 오디오 신호의 spectral magnitude를 예측하게 된다. 이때, azimuth 및 width 파라메타를 조절하여 분리하고자 하는 음원의 spectral magnitude를 제어할 수 있다. 마지막으로, 원음의 위상과 예측된 spectral magnitude에 대해 inverse STFT를 적용하여 객체 오디오를 분리해 낸다. 다음 절에서 제안된 음원분리 기법에 대해서 자세히 기술하도록 한다.

2. 제안된 음원분리 알고리즘

II.1절에서 기술한 바와 같이 M개의 채널로 형성된 비균등 선형 마이크로폰 어레이를 활용하여 입력된 신호는 아래와 같이 표현될 수 있다.

$$\begin{bmatrix} x_1(n) \\ \vdots \\ x_M(n) \end{bmatrix} = \begin{bmatrix} a_1 s(n-\tau_1) \\ \vdots \\ a_M s(n-\tau_M) \end{bmatrix} + \begin{bmatrix} v_1(n) \\ \vdots \\ v_M(n) \end{bmatrix} \quad (12)$$

이때 조향 벡터는 마찬가지로 수식 (4)와 같이 표현가능하다. 다음으로, far-field 환경에서는 객체 오디오 신호의 방향에 따라서 마이크로폰간의 상대적인 시간차가 결정되어지기 때문에 시간차를 추정함으로써 객체 오디오의 방향 또한 추정이 가능하다. 두 개의 마이크로폰 어레이로 구성된 경우 시간차 분석은 다음과 같은 수식으로 표현될 수 있다.

$$\hat{\tau}(k) = \arg\max_{\tau} \left| X_1(k) + W_N^{k\tau(\theta)} X_2(k) \right| \quad (13)$$

여기서, $W_N^{k\tau(\theta)} = \exp(-j2\pi k\tau(\theta)/N)$ 이며, $X_1(k)$ 와 $X_2(k)$ 는 두 개의 마이크로폰으로 획득한 오디오 신호의 k번째 주파수 bin에서의 spectrum을 가리킨다. 그리고 θ는 마이트로폰 어레이의 중앙에서부터 객체 오디오가 놓인 신호의 방향을 의미한다. 즉, 최대값이 되는 τ를 찾음으로써 스테레오 마이크로폰간의 상대적인 시간차를 구한다.

하지만, 수식 (13)을 통해 시간차 분석을 직접적으로 수행하는 대신, 다채널 마이크로폰 어레이에서의 방향에 따른 주파수 영역의 함수, 즉 AF plane을 아래의 수식처럼 정의할 수 있다[9,10,13].

$$AF(k,\theta) = \frac{1}{M} \left| W_N^{k\tau_1(\theta)} X_1(k) + \cdots + W_N^{k\tau_M(\theta)} X_M(k) \right| \quad (14)$$

여기서, $X_i(k)$는 다채널 마이크로폰 어레이에서 i번째 해당하는 마이크로폰으로 획득한 오디오 신호의 k번째 주파수 bin에서의 spectrum을 의미한다. 수식 (14)를 discrete한 θ에 대해 구하기 위해, 본 논문에서는 resolution과 계산량을 고려하여 1° 단위로 θ을 계산하였다. 수식 (14)에서

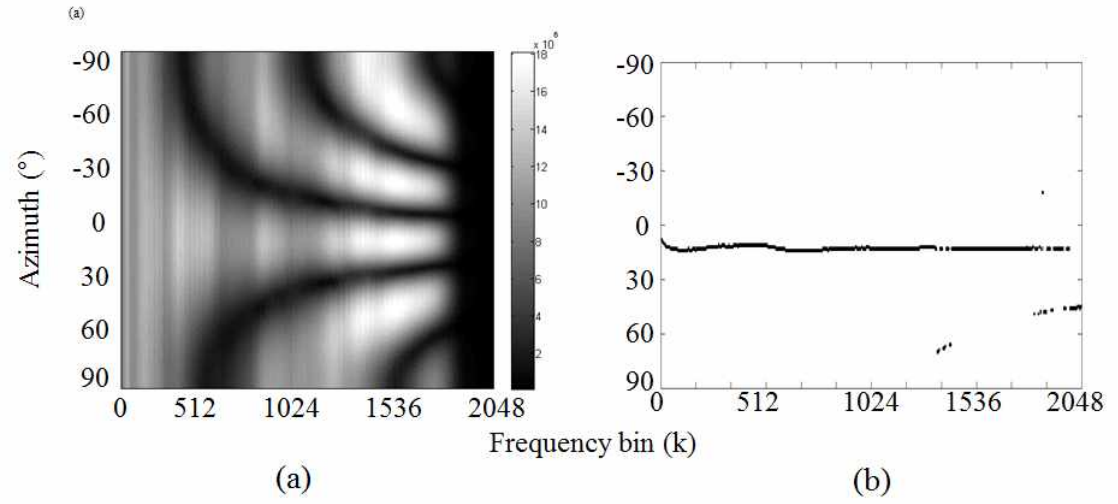


그림 3. 스테레오 마이크로폰 환경에서 15°에 놓인 white noise에 대한 AF plane 분석 결과: (a) $AF(k,\theta)$, (b) $\overline{AF}(k,\theta)$
Fig. 3. Analysis of AF planes for white noise located at 15° in a stereo microphone configuration: (a) $AF(k,\theta)$, (b) $\overline{AF}(k,\theta)$

실제 타겟 신호의 방향이 θ에 근접할수록 $AF(k,\theta)$가 커지게 되기 때문에, $AF(k,\theta)$가 최대가 되는 θ에서 객체 오디오 신호가 있다고 추정할 수 있다. 또한, 타겟 방향의 음원만을 추출하기 위해서 주파수별로 최대가 되는 θ를 제외한 나머지 θ에 대한 $AF(k,\theta)$를 작은 상수 △로 설정함으로써 타겟이 존재하는 방향의 음원이 강조된 AF plane으로 수정할 수 있다.

$$\overline{AF}(k,\theta)=\begin{cases}AF^{\max}(k), & \text{if } AF(k,\theta)=AF^{\max}(k)\\ \triangle, & otherwise\end{cases} \quad (15)$$

여기서, △=0으로 하였으며 $AF^{\max}(k)=\max_{\theta}AF(k,\theta)$이다. 주파수별로 하나의 객체 오디오 신호가 우세하다는 가정 하에, 객체 오디오 신호가 나타나는 방향에만 객체 오디오 신호의 magnitude 값을 설정하고 나머지 방향에는 △로 설정한다.

지금까지 설명한 AF plane을 활용한 음원 위치 추정의 성능을 보이기 위하여, 간격이 5cm인 스테레오 마이크로폰을 통해 48kHz의 표본화율로 입력 받은 white noise를 15°에 time panning[16]하여 음원의 위치 추정 실험을 진행하였다. <그림 3(a)>는 15°에 위치한 white noise를 수식 (14)를 활용하여 계산된 $AF(k,\theta)$를 보여주고, <그림 3(b)>는 수식 (15)로부터 계산된 $\overline{AF}(k,\theta)$를 보여준다. 그림에서 보는 바와 같이, θ=15° 주위에서 $AF(k,\theta)$가 최대값을 갖는 것을 관찰할 수 있으며, $\overline{AF}(k,\theta)$는 θ=15° 주위 이외에는 0으로 표현됨을 알 수 있다. 하지만, 고주파 영역에서는 15° 주위 이외에 다른 영역에 최대값이 형성되는 이유는 고주파로 갈수록 마이크로폰 간격에 비하여 파장이 짧아져 spatial aliasing이 발생하기 때문이며, 또한 저주파 구간에서는 마이크로폰 간격에 비하여 파장이 길어져서 한 파장조차 분석하기 어렵기 때문이다[17].

AF plane을 이용한 음원분리를 위해, azimuth 및 width 파라메타를 활용하는데, 음원분리는 아래의 수식으로 표현될 수 있다.

$$|Y(k)|=\sum_{\theta=d_a-(B/2)}^{d_a+(B/2)}\overline{AF}(k,\theta) \quad (16)$$

여기서, d_a와 B는 분리하고자 하는 음원방향에 대한 azimuth와 width이다. 수식 (16)에서와 같이, 어떤 방위각에 해당하는 신호를 분리할 것인지는 d_a를 통해 결정되며, 얼마만큼의 방위각 넓이로 분리할 것인지는 B를 통해 결정된다. Azimuth d_a와 width B에 대한 개념은 <그림 4>에 나타

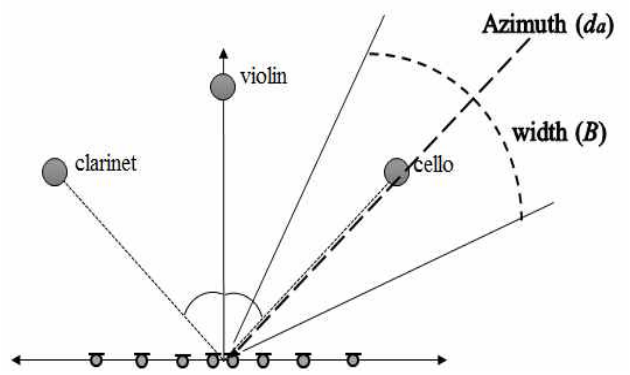


그림 4. Azimuth 및 width 파라메타의 예
Fig. 4. Illustration of azimuth and width parameters

나있다. 그림에서 보는 바와 같이, 마이크로폰 어레이를 기준으로 d_a는 수평 방위각을 가리키며, d_a를 기준으로 얼마만큼의 방위각에 대한 객체 오디오 신호를 가지고 추출할 것인지를 결정하는 것이 B가 된다. 이때 B를 너무 크게 하면, 다른 객체 오디오 신호까지 함께 추출될 수 있으며, B를 너무 작게 하면 분리된 객체 오디오 신호의 음질 열화가 발생하기 쉽다. 그러므로, 적정한 B 파라메타 설정이 중요하다고 할 수 있다. 본 논문에서는 수식 (15)로부터 얻은 $\overline{AF}(k,\theta)$의 histogram을 생성하여 peak가 형성되는 지점을 d_a로 설정하였다. 또한, 객체 오디오 신호 ±15° 주위로는 다른 객체 오디오 신호가 존재하지 않을 거라고 가정하고, B를 30°로 설정하였다. 마지막으로, 수식 (16)으로 획득한 magnitude 성분과 원음의 phase 성분을 가지고 분리된 음원의 spectrum은 다음 수식으로 표현된다.

$$Y(k) = |Y(k)|\exp(j\angle X_l(k)) \tag{17}$$

여기서, $X_l(k)$는 다채널 마이크로폰 어레이에서 설정한 d_a와 가장 가까운 마이크로폰, 즉 l번째 마이크로폰의 으로 획득한 주파수 영역 신호를 가리킨다. 그리고, 4096-point inverse STFT을 적용하고 overlap-add 기법을 통해 객체 오디오 신호를 최종적으로 획득한다.

Ⅳ. 성능 평가

제안된 기법의 성능을 평가하기 위하여 먼저, 방송용 오디오 콘텐츠를 획득하는 환경을 고려하여 <그림 5>에서와 잔향이 있는 소극장에서 마이크로폰 어레이를 배치하여 음원을 수집하였다. 이때, 마이크로폰 어레이의 배치와 객체 오디오의 배치는 <그림 6>과 같다. 그림에서 보는 바와 같이, 8채널의 비균등 선형 마이크로폰을 활용하였고, 가장 가운데에 있는 한 쌍의 마이크로폰 간격은 3cm, 그 다음 쌍의 마이크로폰의 간격은 각각 10cm, 40cm, 120cm가 되도록 배치하였다. 이와 같은 배치는 다양한 균등 및 비균등 마이크로폰 어레이 배치를 통해 음원분리 기술을 적용하였을 경우, 가장 좋은 분리성능을 보일 때의 배치였다. 객체 오디오는 마이크로폰 어레이의 중심으로부터 정면으로 2m

그림 5. 성능 평가를 위한 녹음 환경 및 마이크로폰 배치
Fig. 5. Recording environment and microphone placement for performance evaluation

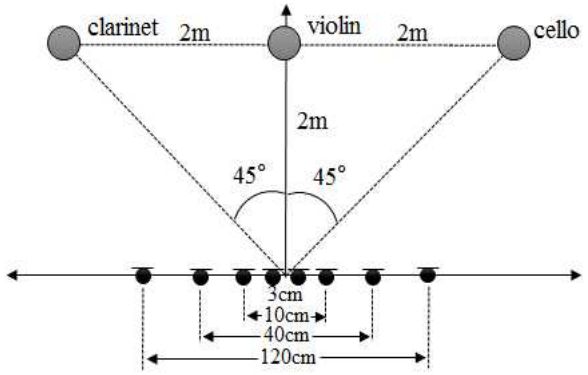


그림 6. 비균등 선형 마이크로폰 어레이 및 객체 오디오의 배치
Fig. 6. Configurations of non-uniform linear microphone array and object-based audio

가 되는 지점에 바이올린을 배치하였고, 바이올린의 위치에서 좌우로 각각 2m가 되는 지점에 클라리넷과 첼로를 배치하였다. 실제 연주 환경을 재현하기 위하여 <그림 7>의 좌측 그림에서와 같이 실제 프로 연주자를 섭외하여 실제 합주하는 소리를 48kHz의 표본화율로 녹음 받았다. 여기서, 실제 녹음된 음원에 대하여 각각의 reference가 될 수 있는 정확한 객체 오디오 획득하기 위한 녹음 방법은 다음과 같다. 우선, 각각의 연주자가 각자의 위치에서 솔로로 연주한 음원을 녹음 받았고, 녹음 받은 음원을 연주가가 연주한 동일한 위치에서 <그림 7>의 우측 그림과 같이 스피커로 동시에 재생하여 합주가 되도록 하고, 이를 다시 재녹음하였다. 이에 따라, 각각의 연주자가 솔로로 연주한 음원을 reference 객체 오디오로 간주하고 성능 평가를 진행하였다.

객관적 성능을 평가하기 위해서 음원분리 기법에서 객관적 척도로 사용되는 SDR (Source-to-Distortion Ratio), SIR (Source-to-Interference Ratio), SAR (Source-to-Artifacts Ratio)를 각각 측정하였다[10]. 여기서, SIR과 SAR은 분리된 객체 오디오 신호가 다른 오디오 신호의 interference를 얼마만큼 적게 받는 지와 음질이 얼마큼 열화되는 지를 각각 나타내고, SDR은 종합적인 distortion을 나타내는 지표이다. 세가지 척도 모두 dB의 단위를 가지며, 높을수록 좋은 성능을 의미한다[11]. 이에 더해, 분리된 음원의 분리정확도도 측정하였다[12]. 분리정확도는 분리된 신호가 실제 reference 음원 중에서 어떤 음원과 가장 유사한 지를 프레임별로 correlation을 측정하고 correlation이 가장 큰 reference 객체와 가장 유사하다고 판단하여 이를 통계적으로 수치화한 것이다[12]. 상대적인 성능 비교를 위해서 MVDR 빔형성기[4]와 ICA 기법[2]으로 처리된 음원의 객관적 성능 수치도 측정하였다.

먼저, <그림 8>은 객체 오디오(클라리넷, 바이올린, 첼로)가 모두 혼합된 입력 신호와, reference 객체 오디오 신호, 제안된 기법으로 분리된 객체 오디오 신호들의 spectrogram을 보여준다. <그림 8(a)>는 클라리넷, 바이올린, 첼로 객체가 모두 혼합된 입력 신호의 spectrogram을, <그림 8(b)>는 reference 첼로 신호, <그림 8(c)>는 분리된 첼로 신호, <그림 8(d)>는 reference 클라리넷 신호, <그림 8(e))

그림 7. 성능 평가를 위한 비균등 선형 마이크로폰 어레이를 활용한 실제 녹음 환경
Fig. 7. Real recording environment using a non-uniform linear microphone array for the performance evaluation

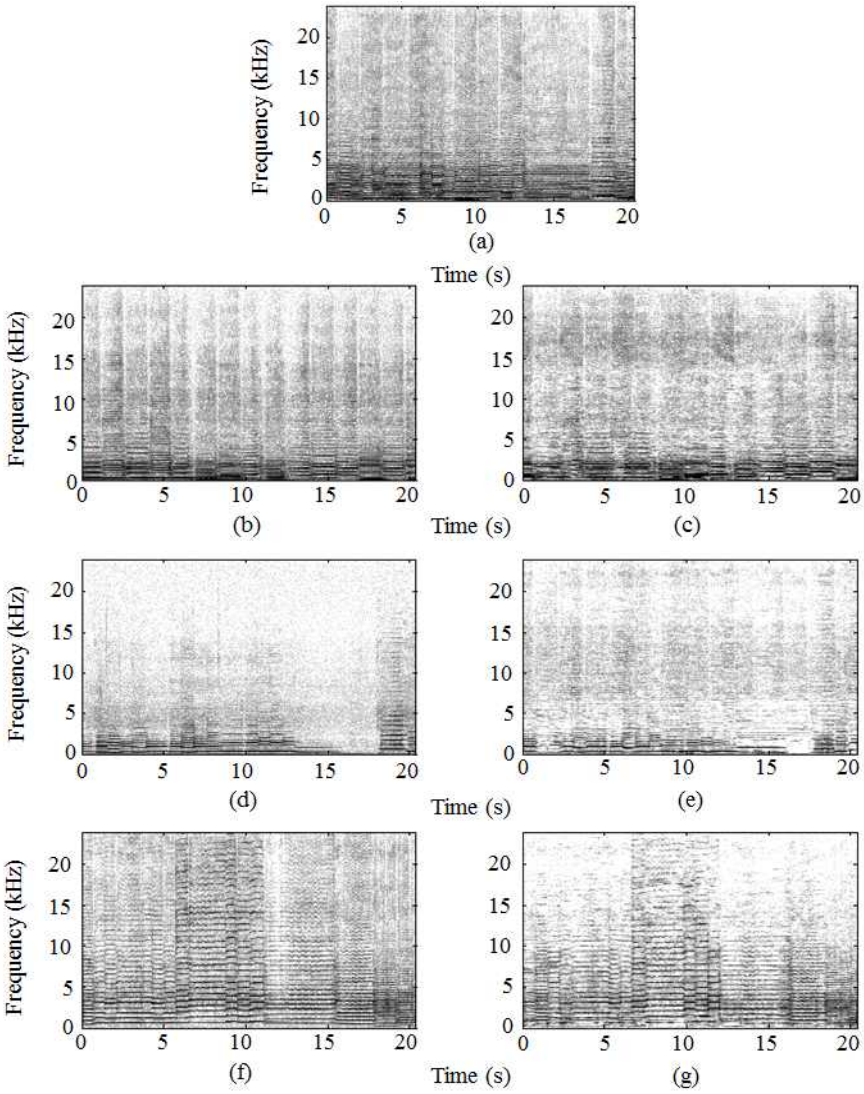


그림 8. 혼합 오디오, reference, 음원분리된 오디오 신호들의 spectrogram 비교: (a) 혼합 오디오 (b) reference 첼로, (c) 음원분리된 첼로, (d) reference 클라리넷, (e) 음원분리된 클라리넷, (f) reference 바이올린, (g) 음원분리된 바이올린

Fig. 8. Comparison of spectrograms of mixtures, references, separated audio signals: (a) mixtures, (b) reference cello, (c) separated cello, (d) reference clarinet, (e) separated clarinet, (f) reference violin, (g) separated violin

는 분리된 클라리넷 신호, <그림 8(f)>는 reference 바이올린 신호, 그리고 <그림 8(g)>는 분리된 바이올린 신호의 spectrogram을 각각 나타낸다. 그림에서 보는 바와 같이, 각각의 분리된 객체 오디오 신호의 spectrogram이 reference 객체 오디오 신호의 spectrogram과 유사한 것을 확인할 수 있었다.

<표 1>은 제안된 기법과 기존의 기법으로 분리된 음원에 대한 SDR, SIR, SAR을 각각 비교하여 보여 준다. 표에서 보는 바와 같이, SDR, SAR, SIR 모두 기존의 MVDR 빔형성기와 ICA 기법에 비하여 제안된 기법이 높은 성능을 보였다. 다음으로, <표 2>는 제안된 기법과 기존의 기법으로 분리된 음원에 대한 분리정확도를 비교하여 보여준

표 1. 제안된 방법과 기존의 방법으로 분리된 음원에 대한 SDR, SIR, SAR 비교
Table 1. Comparison of SDR, SIR, and SAR (dB) of separated audio signals processed by the proposed and conventional methods

Measure	Source	MVDR Method	ICA Method	Proposed Method
SDR	Clarinet	4.82	5.86	5.96
	Violin	3.96	4.83	4.96
	Cello	5.22	6.06	6.59
	Avg.	4.67	5.58	5.84
SAR	Clarinet	7.26	8.04	8.51
	Violin	8.31	7.98	8.52
	Cello	6.89	7.38	7.72
	Avg.	7.49	7.80	8.25
SIR	Clarinet	9.72	9.48	10.36
	Violin	9.91	10.71	10.69
	Cello	10.64	10.53	10.50
	Avg.	10.09	10.24	10.52

표 2. 제안된 방법과 기존의 방법으로 분리된 음원에 대한 분리정확도 비교
Table 2. Comparison of separation accuracy (%) of separated audio signals by the proposed and conventional methods

Source	MVDR Method	ICA Method	Proposed Method
Clarinet	73.6	71.4	83.2
Violin	77.9	70.9	81.7
Cello	78.1	76.4	83.4
Avg.	76.5	72.9	82.8

다. <표 1>과 마찬가지로, 제안된 음원분리 기법의 분리정확도가 기존의 MVDR 빔형성기와 ICA 기법에 비하여 높은 분리정확도를 보였다.

V. 결 론

본 논문에서는 비균등 선형 마이크로폰 어레이 환경에서 고품질의 객체 오디오 콘텐츠 제작을 위한 음원분리 기술을 제안하였다. 제안된 음원분리 기법은 마이크로폰 어레이를 활용하여 채널간의 시간차를 분석하고 AF plane을 생성하였다. 제안된 기법은 주어진 어레이 배치에 따라 채널간의 시간차를 분석하고, 분석된 시간차에 따라 주파수별로 특정 방위각에 위치한 입력 오디오 신호의 magnitude를 예측하였다. 이후, azimuth와 width 파라메타를 조절함으로써 객체 오디오 생성을 위한 음원분리를 수행하였다. 제안된 기법의 성능을 평가하기 위하여 실제 공연이 이루어질 수 있는 소극장에서 실제 연주가가 연주하는 객체 오디오를 녹음 받고, 녹음된 콘텐츠를 활용하여 여러 가지 객관적 성능 지표를 측정하였다. 성능 평가 결과, 제안된 기법이 기존 음원분리 기법들에 비하여 높은 SDR, SAR, SIR를 보였고, 높은 분리정확도를 보였다. 이를 통해 제안된 기법이 기존 기법들의 비하여 상대적으로 적은 음질 왜곡으로 높은 분리 성능을 보이는 것으로 볼 수 있다. 하지만, 고품질 방송용 오디오 콘텐츠 확보의 차원에서 음원분리 기술의 성능에 대한 지표를 함께 연구할 필요가 있다.

참 고 문 헌 (References)

[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] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, “MPEG-H audio－the new standard for universal spatial/3D audio coding,” Journal of the Audio Engineering Society, vol. 62, no. 12, pp. 821-830, Dec. 2014.

[3] S. Makino, T.-W. Lee, and H. Sawada, Blind Speech Separation, Springer, Netherlands, 2007.

[4] A. Hyvärinen, J. Karhunen, and E. Oja, Independent Component Analysis, John Wiley & Sons, Inc., Canada, 2001.

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

[6] H. Cox, R. M. Zeskind, and M. M. Owen, “Robust adaptive beamforming,” IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 35, no. 10, pp. 1365-1375, Oct. 1987.

[7] O. Yilmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” IEEE Transactions on Signal Processing, vol. 52, no. 7, pp. 1830-1847, July 2004.

[8] H. Adel, M. Souad, A. Alaqeeli, and A. Hamid, “Beamforming techniques for multichannel audio signal separation,” International Journal of Digital Content Technology and its Applications, vol. 6, no. 20, pp. 659-667, Nov. 2012.

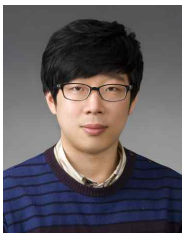
[9] D. Barry, B. Lawlor, and E. Coyle, “Sound source separation: azimuth discrimination and resynthesis,” in Proceedings of International Conference on Digital Audio Effects (DAFX-04), pp. 1-5, Naples, Italy, Oct. 2004.

[10] C. J. Chun and H. K. Kim, “Sound source separation using interaural intensity difference in real environments,” in Proceedings of 135th Audio Engineering Society (AES) Convention, Preprint 8976, New York, NY, Oct. 2013.

[11] E. Vincent, R. Gribonval, and C. Fevotte, “Performance measurement in blind audio source separation,” IEEE Transactions on Audio, Speech

and Language Processing, vol. 14, no. 4, pp. 1462 – 1469, July 2006.
[12] 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.
[13] J. Benesty, J. Chen, and Y. Huang, Microphone Array Signal Processing, Springer, Berlin, Germany, 2008.
[14] M. Brandstein and D. Ward, Microphone Arrays: Signal Processing Techniques and Applications, Springer, Berlin, Germany, 2001.
[15] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," Neural Computation, vol. 9, no. 7, pp. 1483-1492, Oct. 1997.
[16] J. Breebaart, and C. Faller, Spatial Audio Processing: MPEG Surround and Other Applications, John Wiley & Sons, Ltd., Chichester, UK, 2007.
[17] J. Dmochowski, J. Benesty, and S. Affes, "On spatial aliasing in microphone arrays," IEEE Transactions on Signal Processing, vol. 57, no. 4, pp. 1383-1395, Apr. 2009.

저 자 소 개



전 찬 준

- 2009년 한국기술대학교 전자공학과 학사 졸업
- 2011년 광주과학기술원 정보통신공학부 석사 졸업
- 2011년~현재 광주과학기술원 전기전자컴퓨터공학부 박사과정
- ORCID : http://orcid.org/0000-0003-3361-8360
- 주관심분야 : 오디오 신호처리, 3D 오디오

김 홍 국

- 1988년 서울대학교 제어계측공학과 학사 졸업
- 1990년 한국과학기술원 전기 및 전자공학과 석사 졸업
- 1994년 한국과학기술원 전기 및 전자공학과 박사 졸업
- 1990년~1998년 삼성종합기술원 전문연구원
- 1998년~1998년 MMC Technology 선임연구원
- 1998년~2003년 AT&T Labs-Research Senior Member Technical Staff
- 2003년~현재 광주과학기술원 전기전자컴퓨터공학부 교수
- ORCID : http://orcid.org/0000-0002-0105-6693
- 주관심분야 : 음성인식, 음성 및 오디오 신호처리, 3D 오디오