한국음향학회지 제31권 제5호 pp. 317~331 (2012)
The Journal of the Acoustical Society of Korea Vol.31, No.5 (2012)

빔공간-영역 다채널 비음수 행렬 분해 알고리즘을 이용한 음원 분리 기법 Part I: 빔공간-영역 다채널 비음수 행렬 분해 시스템

Audio Source Separation Method Based on Beamspace-domain Multichannel Non-negative Matrix Factorization, Part I: Beamspace-domain Multichannel Non-negative Matrix Factorization system

이석진 · 박상하 · 성굉모

(Seokjin Lee, Sang Ha Park, and Koeng-Mo Sung)

서울대학교 전기컴퓨터공학부 뉴미디어통신공동연구소 음향공학연구실

(접수일자: 2012년 2월 27일; 수정일자: 2012년 4월 19일; 채택일자: 2012년 5월 16일)

초 록: 본 논문에서는 다채널 음향 신호의 음원 분리를 수행하기 위하여, 빔공간-영역에서 다채널 비음수 행렬 분해 기법을 이용하는 음원 분리 시스템을 제안한다. 비음수 행렬 분해(NMF) 기법은 음원 분리에서 최근 널리 쓰이는 알고리즘이며, 특히 최근에는 다채널 비음수 행렬 분해(MC-NMF) 기법으로 발전하여 다채널 음향 신호에 대해서 적용되고 있다. 본 논문에서 제안하는 다채널 비음수 행렬 분해 기법은 빔공간-영역에서 수행되어, 기존의 다채널 비음수 행렬 분해 기법에 비해 좋은 성능을 가진다. 제안되는 비음수 행렬 분해 기법은 SiSEC 2010의 데이터셋을 이용하여 검증되었다.

핵심용어: 다채널 음원 분리, 비음수 행렬 분해, 빔공간 변환
투고분야: 음향 신호 처리 분야(1)

ABSTRACT: In this paper, we develop a multichannel blind source separation algorithm based on a beamspace transform and the multichannel non-negative matrix factorization (NMF) method. The NMF algorithm is a famous algorithm which is used to solve the source separation problems. In this paper, we consider a beamspace-time-frequency domain data model for multichannel NMF method, and enhance the conventional method using a beamspace transform. Our decomposition algorithm is applied to audio source separation, using a dataset from the international Signal Separation Evaluation Campaign 2010 (SiSEC 2010) for evaluation.

Key words: Multichannel audio source separation, non-negative matrix factorization (NMF), beamspace-transform.
ASK subject classification: AcoustIc Signal Processing (1)

I. 서 론

비음수 행렬 분해(NMF: Non-negative Matrix Factorization) 기법은 데이터의 희박 표현을 이용하여 추가적인 정보 없이 하나의 비음수 행렬을 여러 비음수 행렬의 곱으로 나타내는 기법이다. 여기서 희박 표현이란 전체 데이터를 적은 개수의 기저로 대표하여 나타내는 것을 의미한다. 위와 같은 NMF 기법은 2001년 Lee와 Seung에 의해 가중적 갱신법(MU-rule: multiplicative update rule)을 기반으로 하는 NMF 알고리즘 [1]이 제안되면서 영상신호처리와 음향신호처리 분야에 널리 사용되었다.

특히 음향신호처리 분야에서는 음향 신호 분리 문제 [2,3]와 자동 악보 전사 문제 [4-6]에 사용되기 널리 사용되었다. 기본적으로 다루는 데이터가 비음수 행렬 값인 영상신호처리 분야와 달리, 음향신호처리

***Corresponding author:** 이석진 (panind83@gmail.com)
151-742 서울시 관악구 대학동 서울대학교 132동 302호
(전화: 02-880-8427; 팩스: 02-886-0791)

분야에서는 다루는 데이터가 양수와 음수를 모두 포함한다. 따라서, 음향 신호 분리 문제와 자동 악보 전사 문제를 해결하기 위하여, 일반적으로 입력 신호의 스펙트럼의 크기을 취한 후, 이를 분해하는 방법을 사용한다 [5]. 최근에는 단순히 스펙트럼의 크기 값만을 이용하는 방법의 한계를 극복하기 위하여 복소 신호의 NMF를 수행하는 알고리즘 또한 연구되고 있다 [7].

한편, 최근에는 다채널 음향 신호의 음원 분리 문제를 NMF 알고리즘을 이용하여 해결하는 알고리즘이 활발히 연구되고 있다. 기본적으로 이러한 연구들은 각 채널 신호들을 쌓아올려서 하나의 행렬로 만드는 방법이나 [8], 병렬 요소 데이터 모델(PARAFAC: parallel factor data model)을 기반으로 한 비음수 텐서 분해(NTF: non-negative tensor factorization) 알고리즘들을 이용하는 방향 [9]으로 진행되어 왔다. 하지만 위의 알고리즘들은 음향 신호 전달 환경을 전혀 고려하지 않고 일반적인 데이터 분석 툴을 적용하는 방법들이므로 그에 따른 한계를 가진다.

최근에는, Ozerov와 Fevotte에 의해 두 가지 다채널 NMF (MC-NMF: multi-channel NMF) 알고리즘이 개발되었다 [10]. 하나는 기댓값-최대화(EM: expectation-maximization) 기법에 의한 알고리즘이며, 다른 하나는 MU-rule에 의한 알고리즘이다. EM을 기반으로 한 알고리즘이 더욱 좋은 성능을 보여주긴 하지만, 알고리즘에서 필요로 하는 몇 개의 가정들을 만족하지 않으면 크게 성능이 저하되는 문제점을 가지고 있다. MU-rule 알고리즘은 다양한 환경에 강인한 성능을 보이므로, 해당 알고리즘이 더욱 실용적으로 사용될 수 있다.

한편, 음원 분리 문제에서 주목하는 믹싱 환경은 크게 두 가지로 구분된다. 하나는 즉각적 믹싱으로, 각 음원 신호에 단순히 이득만을 적용한 후 합성하는 환경을 의미하며, 다른 하나는 콘볼루티브 믹싱으로, 음원에서 마이크로폰까지 다양한 경로가 존재하는 경우(예를 들어, 반사음이 존재하는 경우)와 같이 음원 신호가 FIR 필터를 거쳐 합성되는 환경을 의미한다. 음원 분리 알고리즘에서 일반적으로 위와 같은 콘볼루티브 믹싱 환경은 전달 FIR 필터에 해당하는 주파수 특성을 곱한 형태로 흔히 모델링되며, 이는 마치 주파수별로 독립적인 전달함수, 즉 믹싱 행렬(mixing matrix)를 가지는 것처럼 모델링되기도 한다 [10].

위의 MC-NMF 알고리즘은 NMF 기반의 다른 다채널 음원 분리 기법에 비해 좋은 성능을 보이지만, 인공적으로 믹싱(mixing)된 신호가 아닌 실제 마이크로폰으로 녹음된 신호의 경우 더욱 개선될 여지가 있다. 기존의 MU-MC-NMF(MU-rule-based MC-NMF) 알고리즘은 입력 신호와 전달 함수의 크기값만을 사용하지만, 신호의 입사각과 같은 공간적인 정보들은 채널 간의 위상 차이에 훨씬 민감하다. 특히 실제 음원이 공간상에 이산적으로 분포하여 있는 콘볼루티브 환경에서는 이러한 문제가 도드라지게 된다. 따라서, 이러한 환경에서는 기존의 MU-MC-NMF 알고리즘이 오동작을 하여 성능이 저하되는 문제가 발생할 수 있다.

따라서, MU-MC-NMF 알고리즘에서 사용하지 않는 채널 간의 위상 값을 사용하게 되면 공간적인 음원 분리에 훨씬 유리할 것임을 예상할 수 있다. 본 논문에서는, 위와 같이 실제 마이크로폰으로 녹음된 환경에서 음원 분리 성능을 높이기 위하여, 빔공간-변환 다채널 NMF(BT-MC-NMF: beamspace-transformed MC-NMF) 알고리즘을 제안한다.

II. 배경 이론

2.1 음향신호 전달환경

2.1.1 시간 영역에서 기술된 신호 전달 환경

음향신호가 전달되는 환경을 명확히 기술하기 위해서, 먼저 마이크로폰 배열이 ULA(uniform line array)의 형태를 이루고 있다고 가정하고, 음향신호는 비교적 원거리, 즉 far-field 환경에서 전달된다고 가정하자. 이 때, m 번째 마이크로폰에 입사되는 음향신호 $x_m(n)$ 은 다음과 같이 시간 지연 함수를 이용하여 모사될 수 있다 [12].

$$x_m(n) = h_m(n) * s(n) \tag{1}$$

$$h_m(n) = \delta(n - mD) \tag{2}$$

여기서 $h_m(n)$ 은 마이크로폰과 음원 간의 전달함수, $s(n)$ 은 음원 신호, $\delta(n)$ 은 디락 델타 함수(Dirac delta function) 함수, 그리고 D는 샘플 지연(sample

delay)를 의미한다. 그림 1에 도시되어 있는 전달 환경을 참고하여 보면, 마이크로폰 사이의 간격 d에 따른 시간 지연 τ 및 샘플 지연 D는 다음과 같은 값을 가짐을 알 수 있다 [12].

$$\tau = \frac{d\sin(\theta_s)}{c} \tag{3}$$

$$D = \frac{\tau}{T_s} = \frac{d\sin(\theta_s)}{c} f_s \tag{4}$$

여기서 θ_s는 음원 신호의 입사각, c는 소리의 전달 속도, 그리고 T_s, f_s는 각각 시스템의 샘플링 주기(sampling period)와 샘플링 주파수(sampling frequency)를 의미한다.

2.1.2 주파수 영역에서 기술된 광대역 신호 전달 환경

한편, (1)의 음향 전달 시스템은 푸리에 변환(Fourier transform)을 이용하면 다음과 같이 주파수 영역에서 기술될 수 있다.

$$X_m(k) = H_m(k)S(k) \tag{5}$$

이 때, 푸리에 변환의 'time shift' 특성에 의해 주파수 영역의 전달 함수는 다음과 같이 구해진다.

$$H_m(k) = e^{-jm\omega_k\tau} \tag{6}$$

여기서 ω_k는 k번째 주파수 빈에 해당하는 각주파수(angular frequency)이다.

2.1.3 공간적 앨리어싱

한편, (6)에서 볼 수 있듯이 far-field 환경에서 입력 신호의 공간적 정보, 즉 입사각에 대한 정보는 입사되는 신호의 위상 차이에 의해 나타나게 된다. 이 때, 그림 2에서 나타나는 것과 같이 고주파에서는 위상 차이로 입사각 정보를 구분할 수 없는 공간적 앨리어싱(spatial aliasing) 문제가 발생하게 된다. 그림 2에서 실선으로 되어 있는 첫 번째 신호(wave 1)와 파선으로 나타나 있는 두번째 신호(wave 2)는 파장의 길이가 같은, 즉 같은 주파수를 가지는 신호이다. 두 신호 모두 마이크로폰과 wave front의 위치가 일치하기 때문에 위상차가 없이 입사되지만, 실제 입사각은 서로 다르다. 이와 같이 위상차로 입사각을 구별할 수 없는 문제를 공간적 앨리어싱 문제라 한다.

ULA(uniform line array)의 경우, 공간적 앨리어싱 문제가 발생하지 않기 위한 조건은 다음과 같다 [12].

$$d \le \frac{\lambda_{\max}}{2} \tag{7}$$

여기서 d는 인접한 마이크로폰 사이의 거리를 의미하며, $\lambda_{\max}$는 가장 높은 목적 주파수에 해당하는 파장을 의미한다. 위의 식을 이용하여 마이크로폰 사이의 거리가 결정된 상황에서 공간적 앨리어싱의 임계 주파수(critical frequency)를 나타내면 다음과 같다.

$$f_c = \frac{2d}{c} \tag{8}$$

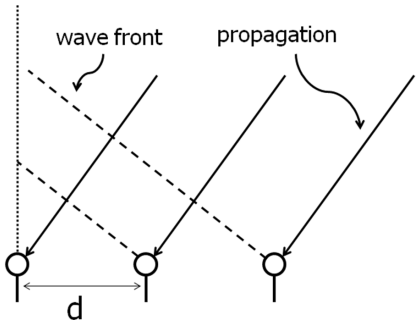

그림 1. Far-field 조건에서의 음향 신호 전달 기술
Fig. 1. Signal propagation description in the far-field condition.

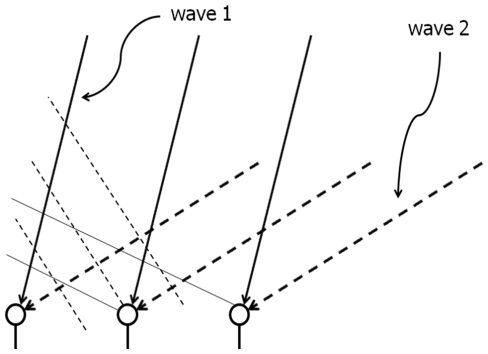

그림 2. 공간적 앨리어싱 문제
Fig. 2. Spatial aliasing problem.

2.2 기존의 다채널 NMF 알고리즘

MU-MC-NMF 알고리즘은 믹싱 환경에 따라 다시 두 가지로 나눠진다. 하나는 콘볼루티브 믹싱을 위한 알고리즘이며, 다른 하나는 즉각적 믹싱을 위한 알고리즘이다. 먼저 콘볼루티브 믹싱 환경에서 다채널 NMF 알고리즘을 도출하기 위하여, 입력 신호를 다음과 같이 모델링한다.

$$x_{i,fn} = \sum_{j=1}^{J} a_{ij,f} s_{j,fn} + b_{i,fn} \tag{9}$$

여기서 $x_{i,fn}$는 국소 푸리에 변환(STFT: short-time Fourier transform)된 i번째 마이크로폰의 입력 신호를, $s_{j,fn}$은 국소 푸리에 변환된 j번째 음원 신호를, 그리고 $b_{i,fn}$은 i 번째 마이크로폰의 잡음신호를 나타낸다. 또한 $a_{ij,f}$는 특정 주파수 빈에서 i번째 마이크로폰과 j번째 음원 간의 전달 함수를 나타내며, 아래 첨자 f와 n은 각각 주파수 빈(frequency bin) 번호와 시간 프레임(time frame) 번호를 나타낸다.

기존의 MU-MC-NMF 알고리즘에서 갱신 수식을 도출하기 위해서, 식 (9)를 이용하여 다음과 같은 입력 신호의 제곱 크기(squared magnitude)를 정의한다 [10].

$$\hat{v}_{i,fn} = \sum_{j} q_{ij,f} \underbrace{\sum_{k \in \mathrm{K}_j} w_{fk} h_{kn}}_{p_{j,fn}} \tag{10}$$

여기서 $\hat{\nu}_{i,fn}$는 $|x_{i,fn}|^2$을 추정한 값을 의미하고, $q_{ij,f}$는 $|a_{ij,f}|^2$을 의미한다. 이 때, MC-NMF 알고리즘은 다음과 같은 비용함수를 최적화하는 것을 목표로 한다.

$$C(\Theta) = \sum_{ifn} d_{IS}\left(\left|x_{i,fn}\right|^2 \middle| \hat{v}_{i,fn}\right) \tag{11}$$

여기서 Θ는 모든 요소(parameter)들의 집합을 의미하며, $d_{IS}(a|b)$는 a와 b 사이의 이타쿠라-사이토 거리(Itakura-Saito divergence)를 의미한다. 위와 같은 이타쿠라-사이토 거리를 최적화하는 해답을 구하는 문제의 경우 일반적으로 해석적(analytic)으로 그 해를 구하기가 어렵고, 특히 위의 NMF 문제와 같이 최적화할 요소가 많은 경우 이를 단번에 구해낼 수가 없다. 따라서, 위의 문제는 MU-rule과 같이 반복적(iterative)인 갱신을 통하여 문제를 풀어나가야 하며, 이 때 갱신식은 다음과 같다 [10].

$$\mathbf{q}_{ij} \leftarrow \mathbf{q}_{ij} \otimes \left\{ \left(\left[\mathbf{V}_i \otimes \left(\mathbf{W}_j \mathbf{H}_j \right) \% \left(\hat{\mathbf{V}}_i \otimes \hat{\mathbf{V}}_i \right) \right] \mathbf{1}_{N\times 1} \right) \otimes \left(\left[\left(\mathbf{W}_j \mathbf{H}_j \right) \% \hat{\mathbf{V}}_i \right] \mathbf{1}_{N\times 1} \right) \right\} \tag{12-a}$$

$$\mathbf{W}_j \leftarrow \mathbf{W}_j \otimes \left\{ \left(\sum_{i=1}^{I} diag\left(\mathbf{q}_{ij}\right) \left[\mathbf{V}_i \% \left(\hat{\mathbf{V}}_i \otimes \hat{\mathbf{V}}_i \right) \right] \mathbf{H}_j^T \right) \% \left(\sum_{i=1}^{I} diag\left(\mathbf{q}_{ij}\right) \left[\mathbf{1}_{F\times N} \% \hat{\mathbf{V}}_i \right] \mathbf{H}_j^T \right) \right\} \tag{12-b}$$

$$\mathbf{H}_j \leftarrow \mathbf{H}_j \otimes \left\{ \left(\sum_{i=1}^{I} \left[diag\left(\mathbf{q}_{ij}\right) \mathbf{W}_j \right]^T \left[\mathbf{V}_i \% \left(\hat{\mathbf{V}}_i \otimes \hat{\mathbf{V}}_i \right) \right] \right) \% \left(\sum_{i=1}^{I} \left[diag\left(\mathbf{q}_{ij}\right) \mathbf{W}_j \right]^T \left[\mathbf{1}_{F\times N} \% \hat{\mathbf{V}}_i \right] \right) \right\} \tag{12-c}$$

여기서 q_{ij}는 $F\times 1$ 길이의 벡터 $\left[q_{ij,f}\right]_f$이고, W_j는 $F\times \mathrm{K}_j$ 크기의 행렬 $\left[\mathrm{W}_{fk}\right]_{fk}$이며, H_j는 $K_j\times N$ 크기의 행렬 $\left[h_{kn}\right]_{kn}$이다. 그리고 연산자 $\otimes$는 하다마드 곱(Hadamard product), 즉 원소별 곱셈(element-wise multiplication)을 의미하며, 연산자 $\%$는 원소별 나눗셈(element-wise division)을 의미한다.

한편, 콘볼루티브 믹싱 환경이 아닌 즉각적 믹싱 환경일 경우 전달 함수가 주파수의 영향을 받지 않으며, 따라서 위의 모델링에서 다음과 같은 관계를 가진다.

$$a_{ij,f} = a_{ij} \tag{13}$$

이 경우, 스펙트럼의 추정된 제곱 크기(estimated squared magnitude)값은 다음과 같이 모델링 된다.

$$\hat{v}_{i,fn} = \sum_{j} q_{ij} \underbrace{\sum_{k \in \mathrm{K}_j} w_{fk} h_{kn}}_{p_{j,fn}} \tag{14}$$

위의 모델링을 가지고 콘볼루티브 믹싱 환경과 동일한 방법론으로 알고리즘을 도출하면, 다음과 같이 q_{ij}를 갱신하는 수식을 얻는다 [10].

$$q_{ij} \leftarrow q_{ij} \left[\frac{\mathbf{1}_{1\times F} \left[\mathbf{V}_i \otimes \left(\mathbf{W}_j \mathbf{H}_j \right) \% \left(\hat{\mathbf{V}}_i \otimes \hat{\mathbf{V}}_i \right) \right] \mathbf{1}_{N\times 1}}{\mathbf{1}_{1\times F} \left[\left(\mathbf{W}_j \mathbf{H}_j \right) \% \hat{\mathbf{V}}_i \right] 1_{N\times 1}} \right] \tag{15}$$

W_j와 H_j의 갱신은 식 (12-b), (12-c)와 유사한 꼴을 가지되, $diag(\mathrm{q}_{ij})$ 대신 q_{ij}를 사용하면 된다 [10].

위에서 살펴본 바와 같이, 기존의 MU-MC-NMF 알고리즘은 전달함수의 크기값만을 이용하여 음원 분리를 수행한다. 하지만 입사각과 같은 공간적 음원 정보는 채널 간의 크기 차이보다는 채널 간의 위상

차이에 훨씬 더 크게 반영된다. 본 논문에서는, 이러한 문제점을 해결하기 위하여 빔공간-영역으로 데이터를 변환한 후, 이에 맞는 MC-NMF 알고리즘을 개발함으로써 채널 간의 위상 차이 정보를 이용하고자 한다.

그림 3에 기존의 채널-시간-주파수 영역의 데이터와 변환된 빔공간-시간-주파수 영역의 데이터가 예시로써 시뮬레이션 되어 비교되었다. 시뮬레이션은 10 cm 간격의 마이크로폰 8개로 이루어진 선배열(line array) 센서와 far-field 전달환경을 가정하여 수행되었다. 그림 3에 도시되어 있는 데이터는 총 200 프레임의 신호가 입사되고 있으며, 전반부의 100 프레임에는 1 kHz의 협대역 신호가 -20° 방향에서 입사되고 있고, 후반부의 200프레임에는 500 Hz와 1 kHz의 성분을 가지는 신호가 40° 방향에서 입사되고 있다. 그림 3의 (a)는 채널-시간-주파수 영역의 크기값 데이터로서, x축은 채널 번호(즉, 마이크로폰 번호), y축은 시간-영역의 프레임 번호, 그리고 z축은 주파수-영역의 빈(bin) 번호를 의미한다. 앞서 언급한 바와 같이 전반부 100 프레임과 후반부 100 프레임의 신호는 서로 다른 방향에서 입사되고 있지만, 그림 3의 (a)에서 보듯이 채널간의 크기값 차이(magnitude difference)는 전반부와 후반부의 차이가 나타나지 않는다. 한편, 그림 3의 (b)는 이를 빔공간-영역(beamspace-domain)으로 변환한 데이터로써, x축은 입사각 정보를 나타낸다. 그림 3의 (a)와 달리, (b)는 전반부 100 프레임과 후반부 100 프레임의 입사각 차이가 뚜렷하게 반영되어 있으며, 이는 빔공간-영역으로 변환하는 과정에서 채널 간 위상 차이(inter-channel phase difference)가 반영되기 때문이다.

그림 3의 (a)와 (b)의 비교를 통해, 빔공간-변환 데이터가 다음과 같은 장점을 가진다는 것을 확인할 수 있다. 첫 번째로는, 채널-시간-주파수 영역의 데이터에 비해 빔공간-시간-주파수의 데이터가 공간상의 음원의 위치 정보를 훨씬 잘 반영한다는 점이다. 이는 그림 3의 (b)에서 전반부와 후반부의 데이터 차이가 확실히 보인다는 것으로 확인할 수 있다. 두 번째로는, 빔공간-시간-주파수의 데이터가 x축 방향으로 훨씬 희박(sparse)한 특성을 가지며, 이는 데이터의 희박 특성을 이용하는 NMF 알고리즘의 특성상 매우 유리하게 작용할 수 있다. 마지막으로, 빔공간-변환 과정이 매우 이상적이어서 주파수에 따른 왜곡이 일어나지 않는다면, 즉 빔공간-영역의 데이터가 신호의 입사각에 따른 값을 정확히 예측할 수 있다면 모든 주파수 성분은 특정 입사각에서 동일하게 큰 값을 가질 것으로 예상할 수 있다. 예를 들어, 실제 광대역 신호가 30° 방향에서 강하게 입사되고 있다면, 이상적인 빔공간-변환이 수행 되었을 때 모든 주파수 성분에서 30°에 해당하는 빔공간(beamspace)에 강한 신호가 관측될 것이다. 이는 곧 주파수-영역에

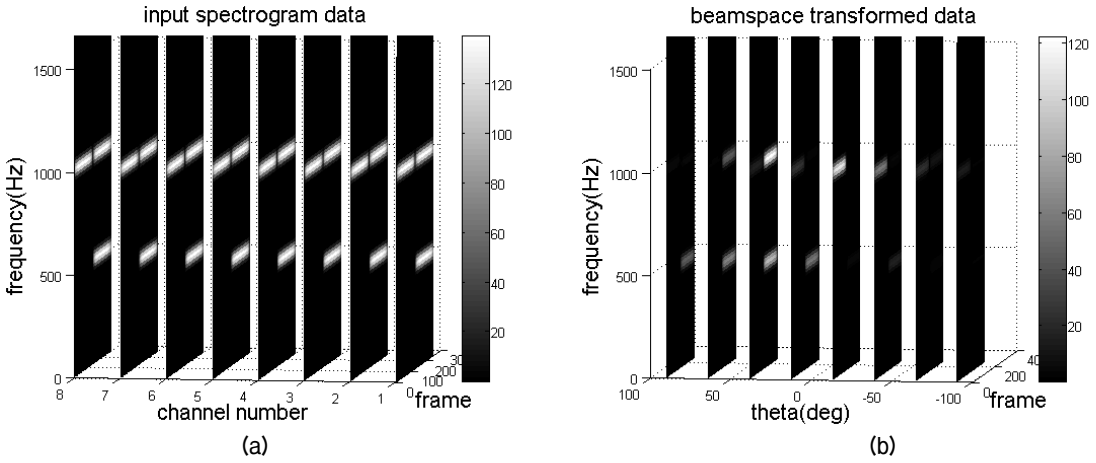

그림 3. 채널-영역 신호와 빔공간-영역 신호 데이터의 비교 예시
(a) 채널-시간-주파수 영역 데이터 (b) 빔공간-시간-주파수 영역 데이터
Fig. 3. An example of comparison between channel-domain and beamspace-domain data.
(a) channel-time-frequency domain data (b) beamspace-time-frequency domain data

서의 물리적 제약(constraint), 즉 모든 주파수-영역 값이 같은 빔공간 응답을 가져야 한다는 것으로 이용될 수 있으므로, 음원 분리 요소들을 추정할 때 이러한 제약 조건을 이용하여 원하지 않는 국소-최소값(local minima)로 수렴하는 것을 제한하고 원하는 해로 수렴시킬 수 있다. 이는 곧 음원 분리 성능의 향상으로 이어진다. 이에 대한 자세한 설명은 다음 장에서 다루어질 것이다.

III. 빔공간-영역에서의 제안하는 신호 모델링

3.1 무향 환경에서의 제안하는 신호 전달 모델링

본 논문에서는 빔공간-영역의 신호 데이터를 그림 4의 (b)와 같이 빔공간-시간-주파수 영역의 3차원 텐서(tensor) 형태로 정의한다. 이 때, 빔공간-영역의 신호 전달 모델링은 다음과 같이 정의된다 [11].

$$\tilde{x}_{\theta,fn} = \sum_{j=1}^{J} \delta\left(\theta - \theta_j\right) s_{j,fn} + b_{\theta,fn} \tag{16}$$

여기서 θ_j은 j번째 음원의 입사각을 의미하며, 빔공간 델타 함수 $\delta(\theta)$은 다음과 같은 값을 가지는 함수로 정의된다.

$$\delta(\theta) = \begin{cases} 1 & \text{if } \theta=0 \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

한편, 식 (16)은 입사각 θ 영역에서 연속적인 함수인 반면, 실제 사용할 수 있는 빔공간-영역 신호는 여러 개의 빔공간 빈(beamspace bin)으로 대표되는 이산 함수이다. 일반적으로, 연속적인 함수 공간에서의 델타 함수는 이산 함수 공간에서의 sinc함수로 모사될 수 있다 [12]. 따라서, 식 (16)의 신호 전달 모델링은 이산 빔공간 영역(discrete beamspace domain)에서 다음과 같이 재정의될 수 있다.

$$\tilde{x}_{m,fn} = \sum_{j=1}^{J} \mathrm{sinc}\left(\theta_m - \theta_j\right) s_{j,fn} + b_{m,fn} \tag{18}$$

여기서 $m(m=1,\cdots,M)$은 빔공간 빈(beamspace bin) 번호를 의미한다.

3.2 잔향 환경으로의 확장 모델링

그림 5와 같이 음원과 마이크로폰 근처에 음원을 반사하는 벽이 하나 존재하는 경우를 고려해보자. 그림 5에 나타나있듯이 반사되어 입사하는 신호를 다른 하나의 가상 음원(image source)에 의한 신호라고 가정하면, 이 경우 식 (18)과 같은 신호 전달 모델링은 다음과 같이 수정될 수 있다 [11].

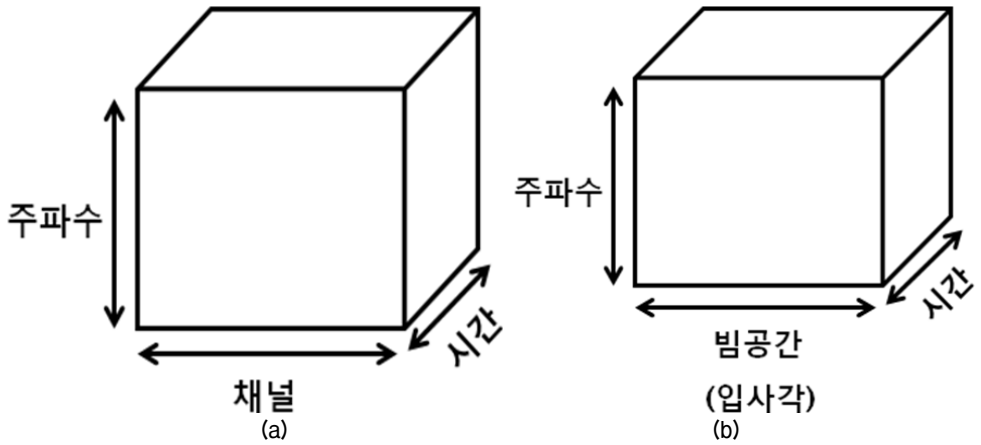

그림 4. 일반적인 채널-영역 데이터와 본 논문에서 제안하는 빔공간-영역 데이터 텐서(tensor) 구조
(a) 채널-영역 데이터 텐서 구조 (b) 빔공간-영역 데이터 텐서 구조
Fig. 4. Tensor structures of conventional channel-domain and proposed beamspace-domain data.
(a) a tensor structure of channel-domain data (b) a tensor structure of beamspace-domain data

$$\tilde{x}_{m,fn} = \sum_{j=1}^{J}\left[\mathrm{sinc}\left(\theta_m - \theta_j\right) + \alpha\mathrm{sinc}\left(\theta_m - \theta_j'\right)\right]s_{j,fn} + b_{m,fn} \quad (19)$$

여기서 $\alpha\,(0<|\alpha|\le 1)$는 복소값을 가지는 상수로, 반사에 의한 신호 크기의 감쇄와 위상차를 모두 반영하는 값이며, θ_j'는 반사된 신호의 입사방향(direction of arrival)을 나타낸다. 만약 신호 환경에 매우 많은 반사음이 존재하는 경우, 즉 잔향 환경인 경우 위의 모델링은 다음과 같이 확장될 수 있다[11].

$$\tilde{x}_{m,fn} = \sum_{j=1}^{J} g_{mj}s_{j,fn} + b_{m,fn} \quad (20)$$

여기서, g_{mj}는 j번째 음원과 m번째 빔공간 빈(beamspace bin)에 의한 임의의 함수를 나타내며, 해당 음원 신호의 빔공간-영역 전달 함수를 의미한다. $|\alpha|$가 일반적으로 0과 1사이의 값을 가지기 때문에, 일반적으로 g_{mj}는 $\theta_m = \theta_j$일 때, 즉 직접음이 입사되는 빔공간 빈에서 가장 큰 값을 가질 것임을 예상할 수 있다.

기존의 NMF 모델(9)에 비교했을 때, 제안하는 신호 전달 모델(20)은 NMF 모델에 사용될 때 몇 가지 장점을 가지고 있다. 먼저, 제안하는 신호 전달 모델은 음원의 공간적 위치에 의한 채널간 위상차(inter-channel phase difference)를 이용할 수 있다. 실제 다수의 마이크로폰으로 far-field에서 입사하는 신호를 기록하는 경우, 음원에서 각 마이크로폰까지의 경로차이에 의해 위상 차이가 존재하게 되며, 상대적으로 크기 차이는 크지 않다. 기존의 NMF 모델 및 그에 따른 알고리즘은 채널 간의 크기차이를 이용하지만, 제안하는 모델링의 경우 빔공간-변환(beamspace transform) 과정에서 유의미한 위상차이를 고려할 수 있다.

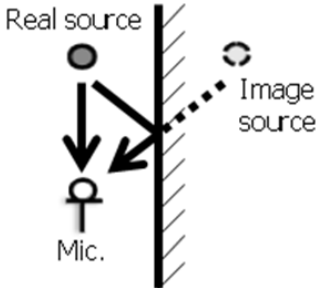

그림 5. 반사 모델
Fig. 5. A reflection model

두 번째로, 제안하는 신호 전달 모델은 모든 주파수 성분이 같은 입사각을 가진다는 물리적 제약 조건을 가진다. 이는 식 (9)의 전달 함수 $a_{i,j,f}$가 주파수에 따라 변하는 값인데 비해 식 (20)의 전달 함수 g_{mj}는 주파수에 무관한 값이라는 것을 통해 알 수 있다. 이러한 물리적 제약 조건은 NMF 알고리즘이 수렴하는 과정에서 무의미한 국소 최저값(local minima)으로 수렴하는 것을 방지할 수 있다.

IV. 제안하는 다채널 NMF 음원 분리 기법

제안하는 음원 분리 기법은 그림 6과 같이 빔공간 변환(beamspace transform), 다채널 비음수 행렬 분해(MC-NMF), 음원 재구성(source reconstruction)으로 구성되어 있다.

일반적으로 협대역 신호의 빔공간 변환은 각 신호의 조향 벡터(steering vector)를 이용하여 수행할 수 있다[13]. 이는 조향벡터가 각 센서 간의 위상 차이를 반영하고 있다는 점에 착안한 것이며, 빔공간 변환은 각 조향 벡터로의 정사영을 구함으로써 수행된다. 이는 마치 시간-주파수 영역에서의 이산 푸리에 변환(Discrete Fourier Transform: DFT)와 같이 작동한다.

구체적인 빔공간 변환 기법을 살펴보면 다음과 같다. 먼저, 협대역 신호 혹은 short-time Fourier transform 된 하나의 주파수 구간(frequency bin)에 대하여, 다음과 같은 빔공간 원형 행렬(beamspace prototype matrix)를 정의한다.

$$\mathbf{W}_{proto,f} = \left[\mathbf{a}_f(\theta_0) \quad \mathbf{a}_f(\theta_1) \quad \cdots \quad \mathbf{a}_f(\theta_M)\right] \quad (21)$$

위의 빔공간 원형 행렬은 $I\times M$크기의 행렬이며, 각 원소를 이루고 있는 $I\times 1$크기의 조향 벡터는 다음과 같이 정의된다.

$$\mathbf{a}_f(\theta) = \left[e^{-j\omega_f(i-1)\left[d_x\sin(\theta)/c\right]}\right]_i \quad i=1,\cdots,I \quad (22)$$

여기서 I는 채널의 개수를 나타내고, d_x는 마이크

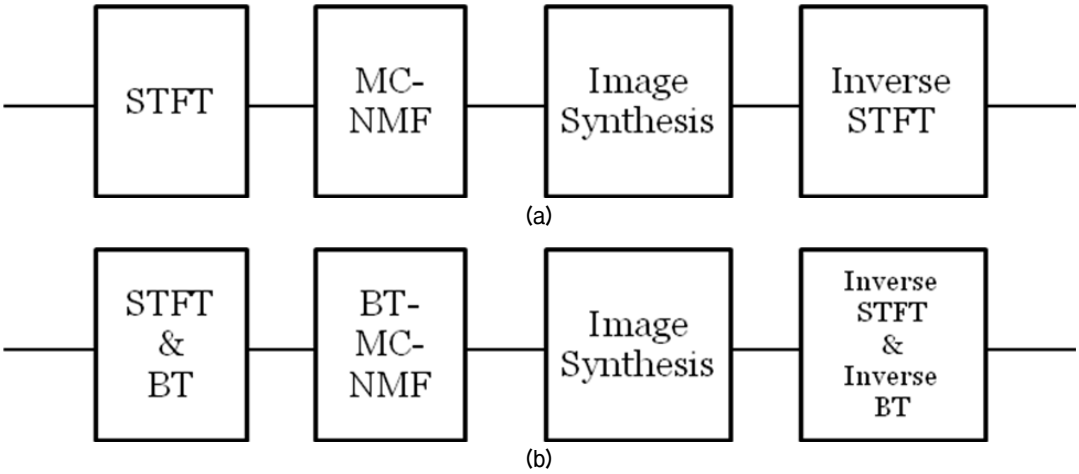

그림 6. 다채널 비음수 행렬 분해 시스템의 블록 다이어그램
(a) Ozerov의 기존 알고리즘 (b) 제안하는 알고리즘
Fig. 6. Block diagrams of multichannel NMF systems.
(a) Conventional MC-NMF system by A. Ozerov (b) Proposed MC-NMF system

로폰 간의 거리를 나타낸다. 조향 벡터가 식 (22)와 같은 형태로 정의되는 것은 식 (5)와 (6)에서 나타난 바와 같이 입사각이 각 센서 신호 간의 위상차이로 나타나기 때문이다. 위의 원형 행렬(prototype matrix)를 이용하여 빔공간 변환을 수행하되, 변환이 sub-unitary property를 갖도록 다음과 같이 보정한다.

$$\mathbf{W}_{BT,f} = \mathbf{W}_{proto,f}\left(\mathbf{W}_{proto,f}^{H}\mathbf{W}_{proto,f}\right)^{-\frac{1}{2}} \tag{23}$$

위와 같이 보정된 빔공간 변환 행렬을 이용하여 다음과 같은 빔공간 변환 작업을 수행한다.

$$\tilde{\mathbf{x}}_{fn} = \mathbf{W}_{BT,f}^{H}\mathbf{x}_{fn} \tag{24}$$

여기서,

$$\tilde{\mathbf{x}}_{fn} = \begin{bmatrix}\tilde{x}_{1,fn} & \cdots & \tilde{x}_{M,fn}\end{bmatrix}^{T} \tag{25}$$

$$\mathbf{x}_{fn} = \begin{bmatrix}x_{1,fn} & \cdots & x_{I,fn}\end{bmatrix}^{T} \tag{26}$$

이다.

4.1 빔공간-영역 다채널 NMF 기법

4.1.1 빔공간-영역 다채널 NMF 알고리즘

위의 빔공간 변환 기법을 이용하면, 3장에서 살펴본 바와 같은 빔공간-영역의 데이터를 얻을 수 있다. 빔공간-영역 데이터의 NMF 기법을 도출하기 위하여, 다음과 같이 이타쿠라-사이토 거리를 이용한 목적함수를 정의한다.

$$C_{NMF}(\Theta) = \sum_{mfn} d_{IS}\left(\left|\tilde{x}_{m,fn}\right|^{2}\middle|\hat{z}_{m,fn}\right) \tag{27}$$

여기서 $\hat{z}_{m,fn}$은 $|\tilde{x}_{m,fn}|^2$의 추정값을 의미하며, 3장의 신호 모델링 하에서 다음과 같이 구성된다.

$$\hat{z}_{m,fn} = \sum_{j} u_{mj} \underbrace{\sum_{k\in \mathrm{K}_j} w_{fk}h_{kn}}_{p_{j,fn}} \tag{28}$$

여기서 $p_{j,fn}$은 j번째 음원의 스펙트럼의 크기(magnitude spectrum)을 나타낸다. 기존의 단일-채널 NMF 기법들에서 이미 밝혀진 바와 같이, 임의의 음향 신호의 magnitude spectrum은 적은 수의 주파수 기저(frequency basis)와 포락선 기저(envelope basis)의 곱으로 표현될 수 있다[2-6]. 여기서 w_{fk}는 k번째 주파수 기저의 f번째 주파수 빈 값을 의미하고, h_{kn}은 k번째 포락선 기저의 n번째 시간 프레임 값을 의미한다. u_{mj}는 위의 제안하는 모델링 (20)의 전달함수 g_{mj}의 크기 제곱을 의미한다.

식 (28)은 Ozerov의 다채널 NMF 모델링 (10)과 같

은 의미를 가지며, 빔공간-영역 데이터에 맞는 변용이라고 볼 수 있다. 식 (28)의 모델링 하에서 식 (27)을 최적화하는 갱신식을 도출하기 위하여, 최적화할 요소의 집합을 $\boldsymbol{\Theta}=\{\mathbf{U},\mathbf{W},\mathbf{H}\}$와 같이 정의하자. 이 때, 집합 $\boldsymbol{\Theta}$에 의한 기울기는 다음과 같이 구해진다.

$$\nabla_{\boldsymbol{\Theta}} C_{NMF}(\boldsymbol{\Theta}) = \sum_{mfn} \left(\nabla_{\boldsymbol{\Theta}} \hat{z}_{m,fn}\right) d'_{IS}\left(\left|x_{m,fn}\right|^2 \middle| \hat{z}_{m,fn}\right) \tag{29}$$

빔공간-영역 다채널 NMF 기법에서 최적화할 요소들은 빔공간 전달 함수인 u_{mj}와, 주파수 기저인 $w_{j,fk}$, 그리고 포락선 기저인 $h_{j,kn}$이다. 따라서, 식 (29)를 위의 3가지 요소에 대하여 다시 표현하면 다음과 같다.

$$\nabla_{u_{mj}} C_{NMF}(\boldsymbol{\Theta}) = \sum_{f=1}^{F}\sum_{n=1}^{N} p_{j,fn} d'_{IS}\left(\left|x_{m,fn}\right|^2 \middle| \hat{z}_{m,fn}\right) \tag{30-a}$$

$$\nabla_{w_{j,fn}} C_{NMF}(\boldsymbol{\Theta}) = \sum_{m=1}^{M}\sum_{n=1}^{N} u_{mj} h_{j,kn} d'_{IS}\left(\left|x_{m,fn}\right|^2 \middle| \hat{z}_{m,fn}\right) \tag{30-b}$$

$$\nabla_{h_{j,fn}} C_{NMF}(\boldsymbol{\Theta}) = \sum_{m=1}^{M}\sum_{f=1}^{F} u_{mj} w_{j,fk} d'_{IS}\left(\left|x_{m,fn}\right|^2 \middle| \hat{z}_{m,fn}\right) \tag{30-c}$$

여기서,

$$d'_{IS}\left(\left|x_{m,fn}\right|^2 \middle| \hat{z}_{m,fn}\right) = \frac{1}{\hat{z}_{m,fn}} - \frac{\left|x_{m,fn}\right|^2}{\hat{z}_{m,fn}^2} \tag{30-d}$$

이다. 위의 수식을 행렬 연산을 이용하여 다시 나타내면 다음과 같다.

$$\nabla_{u_{mj}} C_{NMF}(\boldsymbol{\Theta}) = \mathbf{1}_{1\times F}\left[\mathbf{P}_j \,\%\, \hat{\mathbf{Z}}_m - \mathbf{Z}_m \otimes \mathbf{P}_j \,\%\left(\hat{\mathbf{Z}}_m \otimes \hat{\mathbf{Z}}_m\right)\right]\mathbf{1}_{N\times 1} \tag{31-a}$$

$$\nabla_{\mathbf{W}_j} C_{NMF}(\boldsymbol{\Theta}) = \sum_{m=1}^{M} u_{mj}\left[\left(\hat{\mathbf{Z}}_m - \mathbf{Z}_m\right)\%\left(\hat{\mathbf{Z}}_m \otimes \hat{\mathbf{Z}}_m\right)\right]\mathbf{H}_j^T \tag{31-b}$$

$$\nabla_{\mathbf{H}_j} C_{NMF}(\boldsymbol{\Theta}) = \sum_{m=1}^{M}\left(u_{mj} W_j\right)^T\left[\left(\hat{\mathbf{Z}}_m - \mathbf{Z}_m\right)\%\left(\hat{\mathbf{Z}}_m \otimes \hat{\mathbf{Z}}_m\right)\right] \tag{31-c}$$

한편, 임의의 함수에 대한 기울기 식이 다음과 같이 나타난다고 가정하자.

$$\nabla_{\boldsymbol{\Theta}} C(\boldsymbol{\Theta}) = \nabla^+_{\boldsymbol{\Theta}} C(\boldsymbol{\Theta}) - \nabla^-_{\boldsymbol{\Theta}} C(\boldsymbol{\Theta}) \tag{32}$$

가중적 갱신법(MU-rule)에 의하면, 최적화될 요소는 다음과 같은 갱신식으로 최적화가 가능하다 [1-2].

$$\boldsymbol{\Theta} \leftarrow \boldsymbol{\Theta} \otimes \left[\nabla^-_{\boldsymbol{\Theta}} C(\boldsymbol{\Theta}) \,\%\, \nabla^+_{\boldsymbol{\Theta}} C(\boldsymbol{\Theta})\right] \tag{33}$$

위와 같은 가중적 갱신법을 사용하면, 위의 (31) 식들은 아래와 같은 갱신식으로 변환될 수 있다.

$$u_{mj} \leftarrow u_{mj}\left[\frac{\mathbf{1}_{1\times F}\left[\mathbf{Z}_m \otimes \left(\mathbf{W}_j\mathbf{H}_j\right)\%\left(\hat{\mathbf{Z}}_m \otimes \hat{\mathbf{Z}}_m\right)\right]\mathbf{1}_{N\times 1}}{\mathbf{1}_{1\times F}\left[\left(\mathbf{W}_j\mathbf{H}_j\right)\%\hat{\mathbf{Z}}_m\right]\mathbf{1}_{N\times 1}}\right] \tag{34-a}$$

$$\mathbf{W}_j \leftarrow \mathbf{W}_j \otimes \left\{\left(\sum_{m=1}^{M} u_{mj}\left[\mathbf{Z}_m \%\left(\hat{\mathbf{Z}}_m \otimes \hat{\mathbf{Z}}_m\right)\right]\mathbf{H}_j^T\right)\%\left(\sum_{m=1}^{M} u_{mj}\left[\mathbf{1}_{F\times N} \%\, \hat{\mathbf{Z}}_m\right]\mathbf{H}_j^T\right)\right\} \tag{34-b}$$

$$\mathbf{H}_j \leftarrow \mathbf{H}_j \otimes \left\{\left(\sum_{m=1}^{M}\left[u_{mj}\mathbf{W}_j\right]^T\left[\mathbf{Z}_m \%\left(\hat{\mathbf{Z}}_m \otimes \hat{\mathbf{Z}}_m\right)\right]\right)\%\left(\sum_{m=1}^{M}\left[u_{mj}\mathbf{W}_j\right]^T\left[\mathbf{1}_{F\times N} \%\, \hat{\mathbf{Z}}_m\right]\right)\right\} \tag{34-c}$$

4.1.2 공간적 앨리어싱에 대한 보상

앞서 2장에서 언급하였듯이, 입사각을 이용하여 전달 신호를 분석할 때에는 공간적 앨리어싱 문제가 발생할 수 있다. 빔공간-영역의 전달함수 u_{mj}를 갱신할 때 위와 같은 공간적 앨리어싱 문제를 회피하기 위하여, 식 (34-a)를 다음과 같이 주파수-선택적인 갱신식으로 변경하도록 한다.

$$u_{mj} \leftarrow u_{mj}\left[\frac{\mathbf{1}_{1\times F}\left[\mathbf{Z}_m \otimes \left(\mathbf{W}_j\mathbf{H}_j\right)\%\left(\hat{\mathbf{Z}}_m \otimes \hat{\mathbf{Z}}_m\right)\otimes \mathbf{P}\right]\mathbf{1}_{N\times 1}}{\mathbf{1}_{1\times F}\left[\left(\mathbf{W}_j\mathbf{H}_j\right)\%\hat{\mathbf{Z}}_m \otimes \mathbf{P}\right]\mathbf{1}_{N\times 1}}\right] \tag{35}$$

여기서 $K\times N$크기의 2진 마스크 행렬(binary mask matrix) P는 다음의 2진값을 원소로 가지는 행렬이다.

$$P_{k,n} = \begin{cases} 1 & f_k \le f_c \\ 0 & \text{otherwise} \end{cases} \tag{36}$$

4.2 빔공간 역변환과 음원 재구성

위의 MC-NMF 알고리즘은 음원 신호의 크기는 나타낼 수 있으나, 위상(phase)정보는 복원할 수 없다. 이는 NMF를 이용한 음원 분리 기법의 대부분이 가지고 있는 문제이며, 일반적으로 위상 정보를 최소 평균자승오차(MMSE: minimum mean square error) 관점에서 다음과 같이 추정한다 [10].

$$\tilde{s}_{j,fn}^{(m)im} = \frac{u_{mj} p_{j,fn}}{\hat{z}_{m,fn}} \tilde{x}_{m,fn} \tag{37}$$

여기서 $\hat{s}_{j,fn}^{(m)im}$는 j번째 음원의 m번째 빔공간 빈에 해당하는 추정된 음원 신호를 의미한다.

위와 같이 재구성(reconstruction)된 신호는 빔공간-영역의 신호이므로, 이를 다시 채널-영역으로 되돌리는 역변환(inverse transform)과정이 필요하다. 위의 빔공간 변환 행렬이 sub-unitary property 특성을 가지도록 설계되었으므로, 역변환 과정은 다음의 과정을 통하여 수행될 수 있다.

$$\hat{\mathbf{s}}_{j,fn}^{im} = \mathbf{W}_{BT,f}\tilde{\mathbf{s}}_{j,fn}^{im} \tag{38}$$

여기서,

$$\hat{\mathbf{s}}_{j,fn}^{im} = \left[\hat{s}_{j,fn}^{(1)im} \quad \cdots \quad \hat{s}_{j,fn}^{(I)im}\right] \tag{39-a}$$

$$\tilde{\mathbf{s}}_{j,fn}^{im} = \left[\tilde{s}_{j,fn}^{(1)im} \quad \cdots \quad \tilde{s}_{j,fn}^{(M)im}\right] \tag{39-b}$$

이다.

V. 빔공간-영역 다채널 NMF 기법 성능평가

5.1 성능평가 데이터

제안하는 빔공간-영역 다채널 NMF 기법의 성능을 평가하기 위하여, 2010 Signal Separation Evaluation Campaign(SiSEC 2010)의 "Source separation in the presence of real-world background noise task" 데이터셋을 사용하였다 [14]. 본 데이터셋은 3개의 음성신호원(speech source)과 강한 잡음 신호(background noise)가 존재하는 환경에서 녹음되었으며, 2-채널 데이터와 4-채널 데이터의 2가지 종류로 구성되어 있다, 2-채널 데이터의 경우 2개의 마이크로폰을 사용하여 녹음되었고, 4-채널 데이터의 경우 4개의 마이크로폰으로 구성된 ULA(uniform line array)를 이용하여 녹음되었다. 녹음 환경에 대한 공간적 정보는 그림 7에서 확인할 수 있다.

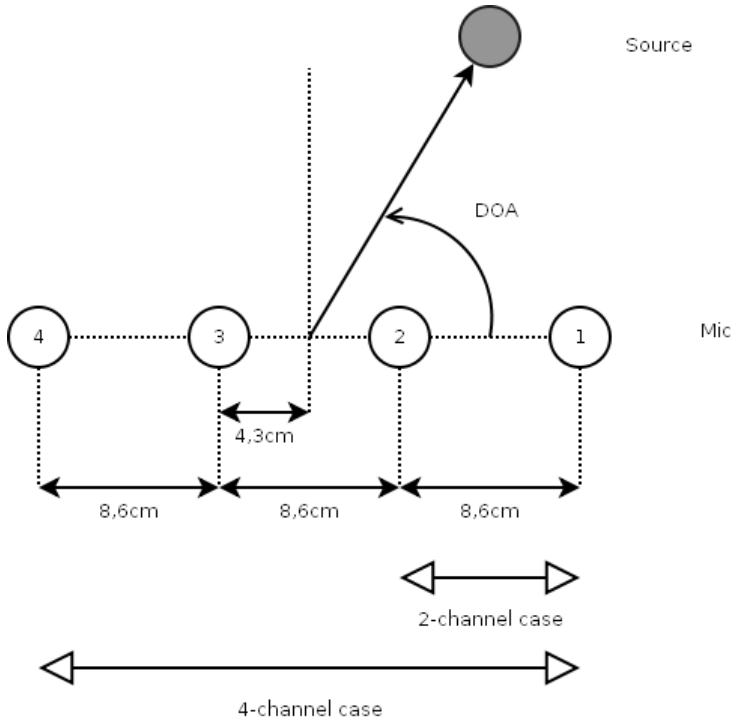

그림 7. SiSCE 2010 "Source separation in the presence of real-world background noise task"의 녹음 환경 [14]
Fig. 7. A recording environment of SiSEC 2010 "Source separation in the presence of real-world background noise task" [14].

표 1. 2-채널 데이터셋의 데이터 세부 특성
Table 1. Data specification of the 2-channel data set.

혼합신호 구분	총 음성음원 개수	남성 음성 개수	여성 음성 개수	잡음 종류	잡음의 등방성
1	3	2	1	Cafeteria	높음
2	3	1	2	Cafeteria	높음
3	3	2	1	Cafeteria	낮음
4	3	2	1	Cafeteria	낮음
5	3	2	1	Square	높음
6	3	1	2	Square	높음
7	3	1	2	Square	낮음
8	3	3	0	Square	낮음
9	3	2	1	Subway	높음
10	3	1	2	Subway	높음

표 2. 4-채널 데이터셋의 데이터 세부 특성
Table 2. Data specification of the 4-channel data set.

혼합신호 구분	총 음성음원 개수	남성 음성 개수	여성 음성 개수	잡음 종류	잡음의 등방성
1	3	1	2	Cafeteria	높음
2	3	1	2	Cafeteria	높음
3	3	1	2	Cafeteria	낮음
4	3	1	2	Cafeteria	낮음
5	3	1	2	Square	높음
6	3	2	1	Square	높음
7	3	1	2	Square	낮음
8	3	1	2	Square	낮음
9	3	1	2	Subway	높음
10	3	1	2	Subway	높음

각 데이터셋은 서로 다른 음원과 잡음 조합을 가진 10개의 혼합신호(mixed signal)으로 구성되어 있으며, 각 혼합 신호를 3개의 음원 신호로 분리하는 것이 주어진 task의 목표이다. 각 데이터셋에 포함된 실험 데이터들의 상세 정보를 비교한 내용이 표 1과 표 2에 나타나있다. 각 음원에 사용된 음성 음원은 다양한 음원들 중 선택되어 사용된 것이며, 특히 표 1의 3번 신호와 4번 신호의 경우와 같이 사용된 음원의 개수가 같은 경우도 있으나, 이 경우 서로 다른 특성의 음원이 사용되었기 때문에 혼합 신호 또한 다른 특성을 가진다.

5.2 알고리즘 파라미터 및 성능 평가 지수

빔공간-영역 다채널 NMF 기법을 수행하기 위하여, 본 실험에서는 50 % 중첩(overlap)되는 1024 길이의 Hamming 윈도우를 사용하여 2048개의 DFT-point로 국소 푸리에 변환(STFT: short-time Fourier transform)을 수행하였다.

제안하는 알고리즘 및 비교 대상의 NMF 알고리즘에 필요한 각 음원 당 기저(basis) 개수 K_j는 40개로 설정하였다. 각 NMF 알고리즘들은 200번의 반복을 통해 갱신이 수행되었으며, 특히 EM알고리즘을 기반으로 한 기법들은 500번의 반복을 통해 갱신이 수행되었다. 빔공간 빈(beamspace bin)의 개수 M은 마이크로폰의 개수 I와 동일하게 설정되었다.

성능 평가 지수로는 Vincent등에 의해 개발된 SDR, SIR, SAR, ISR의 4가지 지표를 하였으며, 이는 음원 분리 기법(source separation method)를 평가할 때 널리 사용되는 지표들이다 [15]. SDR(Signal-to-Distortion Ratio)는 전체적인 에러에 대한 성능 지표이고, SIR

(Source-to-Interference Ratio)는 분리된 음원과 섞여있는 간섭 신호(interference) 신호의 에너지 비를 나타낸다. SAR(Source-to-Artifacts Ratio)는 분리된 음원과 그 자체의 에러(artifacts)의 에너지 비를 나타내고, ISR(source-Image-to-Spatial-distortion Ratio)는 공간적인 에러에 대한 성능 지표이다.

위의 성능 평가 지수는 다음과 같이 정의된다. 먼저 추정된 음원 이미지(sound source image) 신호를 식(40)과 같이 나타낸다 [15].

$$\hat{s}_{ij}^{img}(t) = s_{ij}^{img}(t) + e_{ij}^{spat}(t) + e_{ij}^{interf}(t) + e_{ij}^{artif}(t) \quad (40)$$

여기서 $s_{ij}^{img}(t)$는 i번째 채널, j번째 음원의 실제 이미지 신호를, $e_{ij}^{spat}(t)$, $e_{ij}^{interf}(t)$, $e_{ij}^{artif}(t)$ 신호는 각각 공간적 에러 요소, 간섭 신호 에러 요소, 결함 에러 요소를 나타낸다. 이 때, 각각의 성능 지표는 다음과 같이 정의된다.

$$ISR_j = 10\log_{10}\frac{\sum_{i=1}^{I}\sum_t s_{ij}^{img}(t)^2}{\sum_{i=1}^{I}\sum_t e_{ij}^{spat}(t)^2} \quad (41)$$

$$SIR_j = 10\log_{10}\frac{\sum_{i=1}^{I}\sum_t \left(s_{ij}^{img}(t) + e_{ij}^{spat}(t)\right)^2}{\sum_{i=1}^{I}\sum_t e_{ij}^{interf}(t)^2} \quad (42)$$

$$SAR_j = 10\log_{10}\frac{\sum_{i=1}^{I}\sum_t \left(s_{ij}^{img}(t) + e_{ij}^{spat}(t) + e_{ij}^{interf}(t)\right)^2}{\sum_{i=1}^{I}\sum_t e_{ij}^{artif}(t)^2} \quad (43)$$

$$SDR_j = 10\log_{10}\frac{\sum_{i=1}^{I}\sum_t s_{ij}^{img}(t)^2}{\sum_{i=1}^{I}\sum_t \left(e_{ij}^{spat}(t)^2 + e_{ij}^{interf}(t) + e_{ij}^{artif}(t)\right)^2} \quad (44)$$

각 에러 요소를 추정하는 과정은 [15]에 더욱 상세히 기술되어 있다. 위의 모든 성능 지표는 값이 클수록 좋은 성능을 의미한다. 또한 SDR이 종합적인 성능을 의미하기는 하지만, 각 에러 요소들(공간적 에러 요소, 간섭 신호 에러 요소, 결함 에러 요소)이 독립적으로 인지되어 성능에 영향을 주므로 [15], SIR, SAR, ISR을 독립적으로 비교하는 것이 유의미하며, SDR의 경우 각 요소에 대한 성능 경향이 다를 경우 종합적 성능에 대한 비교 지표로서 해석될 수 있다.

본 논문에서 사용된 성능 평가를 위한 MATLAB 코드는 SiSEC 2010 홈페이지에서 제공되는 것을 사용하였다 [14]. 또한, 한 개의 실험 데이터에 대하여 각 성능지표가 음원 개수만큼 존재하게 되는데(본 실험에서는 3개), 본 논문에서는 성능 비교상의 편의를 위하여 한 개의 실험 데이터에서 각 음원에 대한 성능 지표들의 평균을 취한 하나의 데이터를 도시하였다.

5.3 성능 평가 결과

그림 8과 그림 9는 각각 2-채널 데이터, 4-채널 데이터에 대한 결과를 나타내고 있다. 그림 8에서 볼 수 있듯이, 전반적으로 제안하는 빔공간-영역 다채널 NMF 알고리즘의 성능이 높게 나타나고 있으며, 특히 SIR과 ISR 성능에서 더욱 좋은 성능을 나타냄을 확인할 수 있다. 특히, 기존 다채널 NMF 알고리즘의 경우 간혹 신호를 음원 별로 분리하지 못하고 한 음원 내에서 주파수 영역 별로 분리하는 경우(예를 들어, 기본주파수와 배음 성분을 각각 따로 분리하는 경우)가 발생하는 데 비해, 제안하는 다채널 NMF 알고리즘의 경우 해당 문제점을 보이지 않았다. 그림 8의 7번 실험과 같은 경우가 이에 해당하며, 기존 다채널 NMF 알고리즘의 성능이 SIR 측면에서 크게 저하되었으나 제안하는 알고리즘은 저하되지 않았음을 확인할 수 있다.

그림 9의 4-채널 데이터 결과에서 보듯이, 제안하는 알고리즘은 4-채널 데이터에서도 전반적으로 가장 좋은 성능을 보여주고 있으며, 특히 SDR, SIR, ISR 성능에서 좋은 성능을 보여준다. 또한, 그림 9의 4-채널 데이터의 결과를 보면 그림 8의 2-채널 데이터의 결과에 비해 제안하는 알고리즘의 성능이 향상된 것을 볼 수 있으며, 이를 통하여 채널 개수, 즉 마이크로폰 개수가 늘어날 수록 제안하는 알고리즘이 향상될 것을 예상할 수 있다. 이는 마이크로폰 개수가 늘어날수록 빔공간-영역의 분해능이 증가하는 것에 기인한다.

2-채널 데이터와 4-채널 데이터셋 모두 9,10번 데이터의 성능이 좋지 않은데, 이는 다른 신호에 비해 입력 SNR이 좋지 않기 때문이며, 이에 따라 기존의 알고리즘과 제안하는 알고리즘 모두 성능이 좋지 않음을 확인할 수 있다. 이 때에도 제안하는 알고리즘은 기존 알고리즘에 비해 SIR 성능은 좋은 모습을 보이나, SAR 측면에서 약점이 있음을 확인할 수 있다.

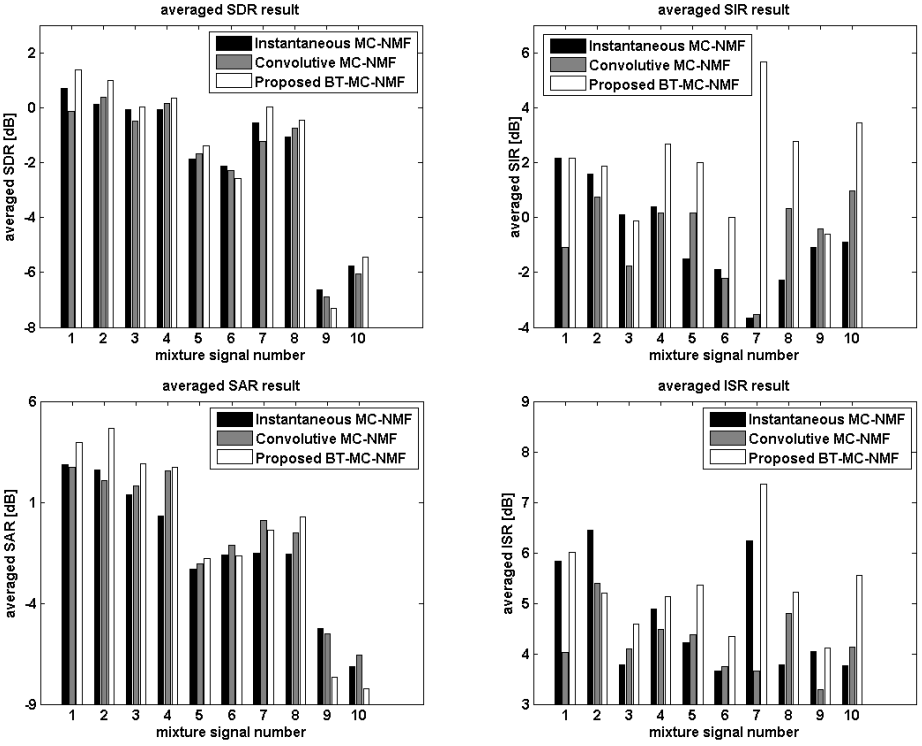

그림 8. 2-채널 신호의 성능 평가 결과
Fig. 8. Evaluation result using 2-channel input signals.

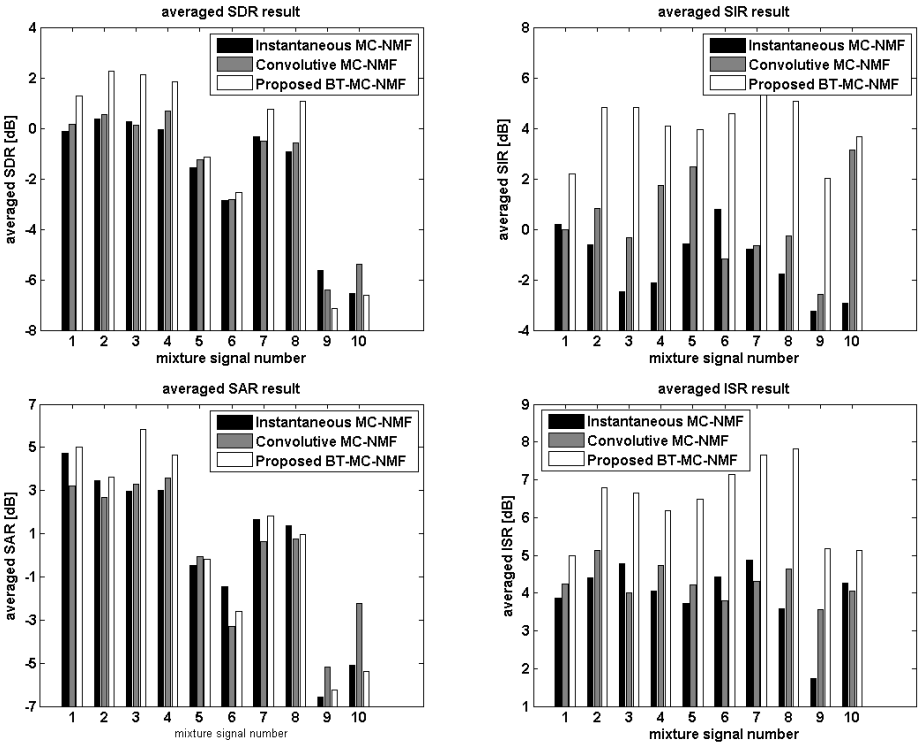

그림 9. 4-채널 신호의 성능 평가 결과
Fig. 9. Evaluation result using 4-channel input signals.

VI. 결 론

본 논문에서는, Ozerov의 다채널 NMF 알고리즘을 기반으로 하되, 실제 음원의 전달 환경을 고려한 개선된 다채널 NMF 알고리즘을 제안하였다. 본 논문에서 제안하는 다채널 NMF 알고리즘은, NMF 알고리즘이 음원의 크기 정보를 이용하지만, 실제 음원이 전달되는 환경에서 각 채널간 음원의 크기는 상대적으로 의미가 없다는 점에 착안하여 제안되었다. 제안된 알고리즘은 위와 같은 문제를 해결하기 위하여 단순한 채널-영역의 다채널 NMF 알고리즘 대신 빔공간-영역의 NMF 알고리즘을 사용하도록 하였다.

빔공간-영역의 다채널 NMF 알고리즘을 도출하기 위하여, 먼저 무향 환경에서의 음원 전달 모델링을 가정한 후, 이를 잔향 환경으로 확장하여 빔공간-영역의 NMF 모델링을 제안하였다. 제안된 빔공간-영역 NMF 모델링 하에서, 추정할 요소들을 최적화하는 NMF 갱신식들을 도출하였다. 이를 통하여, 빔공간-변환, 빔공간-영역 다채널 NMF, 그리고 빔공간-역변환 및 음원 재구성 모듈로 구성되는 빔공간-영역 다채널 NMF 음원 분리 시스템을 제안하였다.

제안된 음원 분리 시스템은 음원 분리 기법의 성능 평가에 사용되는 SiSEC 2010의 데이터셋을 이용하여 검증되었으며, 이를 통하여 기존의 알고리즘에 비해 좋은 성능을 보임을 확인하였다. 특히, 제안하는 알고리즘은 채널 개수가 더 많은 상황에서 효과적으로 동작할 수 있는 가능성을 가짐을 확인할 수 있었다.

참고문헌

1. D. D. Lee and H. S. Seung, “Learning the parts of objects with non-negative matrix factorization,” *Nature*, vol. 401, pp. 789-791, 1999.
2. T. Virtanen, “Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1066-1074, 2007.
3. B. Wang, M. D. Plumbley, “Musical audio stream separation by non-negative matrix factorization,” *in Proc. DMRN Summer Conf.*, 2005.
4. P. Smaragadis and J. C. Brown, “Non-negative matrix factorization for polyphonic music transcription,” *in Proc. IEEE Workshop Applicat. Signal Process. Audio Acout. (WASPAA)*, pp. 177-180, 2003.
5. E. Vincent, N. Berlin, R. Badeau, “Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription,” *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2008)*, pp. 109-112, 2008.
6. 박상하, 이석진, 성굉모, “비음수 행렬 분해(NMF)를 이용한 악보 전사,” *한국음향학회지*, 제29권, 제2호, pp. 102-110, 2010.
7. H. Kameoka, No no, K. Kashino, S. Sagayama, “Complex NMF: A new sparse representation for acoustic signals,” Int. Conf. on Acoustics, Speech and Signal Processing(ICASSP 2009), pp. 3437-3440, April 2009.
8. R. M. Parry and I. A. Essa, “Estimating the spatial position of spectral components in audio,” *in Proc. 6th Int. Conf. Ind. Compon. Anal. Blind Signal Separation (ICA’06)*, pp. 666-673, 2006.
9. D. FitzGerald, M. Cranitch, and E. Coyle, “Non-negative tensor factorization for sound source separation,” *in Proc. Irish Signals Syst. Conf.*, pp. 8-12, 2005.
10. A. Ozerov, C. Févotte, “Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio source Separation,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 550-563, 2010.
11. S. Lee, S. H. Park, K. -M. Sung. “Beamspace-domain Multichannel Nonnegative Matrix Factorization for Audio Source Separation,” *IEEE Signal Processing Letters*, vol. 19, no. 1, pp. 43-46, 2012.
12. C. L. Koh, “Broadband adaptive beamforming with low complexity and frequency invariant response,” *Ph. D. Dissertation*, University of Southampton, 2009.
13. M. D. Zoltowski, G. M. Kautz, S. D. Silverstein, “Beamspace Root-MUSIC,” IEEE Trans. Signal Processing, vol. 41, no. 1, pp. 344-364, 1993.
14. *Signal Separation Evaluation Campaign 2010 (SiSEC 2010),* http://www.sisec.wiki.irisa.fr, 2010.
15. E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, “First stereo audio source separation evaluation campaign: Data, algorithms and results,” *in Proc. Int. Conf. Ind. Compon. Anal. Blind Source Separation (ICA’07)*, pp. 552-559, 2007.

저자 약력

▸ 이 석 진(Seokjin Lee)

2006년 8월: 서울대학교 공과대학 전기·컴퓨터공학부(학사)
2008년 8월: 서울대학교 공과대학 전기·컴퓨터공학부(석사)
2008년 9월 ~ 현재: 서울대학교 공과대학 전기·컴퓨터공학부 박사과정
〈관심분야〉 음원 분리, 음향신호처리, 적응신호처리

▸ 박 상 하(Sang Ha Park)

2005년 2월: 서울대학교 음악대학 기악과(학사)
2008년 2월: 서울대학교 공과대학 전기·컴퓨터공학부(석사)
2008년 3월 ~ 현재 : 서울대학교 공과대학 전기·컴퓨터공학부 박사과정
〈관심분야〉 음원 분리, 악보 전사, 음향신호처리

▸ 성 굉 모(Koeng-Mo Sung)

1971년: 서울대학교 전자공학과(학사)
1973년: 독일 아헨공대 Vordiplom
1977년: 독일 아헨공대 전자통신공학 Dipl.-Ing.
1982년: 독일 아헨공대 음향공학 Dr.-Ing.
1977년 ~ 1983년: 독일 아헨공대 음향공학연구소 연구원
1983년 ~ 현재: 서울대학교 공과대학 전기·컴퓨터공학부 교수