Journal of The Institute of Electronics and Information Engineers Vol.54, NO.2, February 2017 https://doi.org/10.5573/ieie.2017.54.2.123
ISSN 2287-5026(Print) / ISSN 2288-159X(Online)

논문 2017-54-2-16

이중 마이크로폰을 이용한 비음수 행렬분해 기반 다중음원 도래각 예측

(Nonnegative Matrix Factorization Based Direction-of-Arrival Estimation of Multiple Sound Sources Using Dual Microphone Array)

전 광 명*, 김 홍 국**, 유 승 우***

(Kwang Myung Jeon, Hong Kook Kim©, and Seung Woo Yu)

요 약

본 논문에서는 이중 마이크로폰 배열을 이용하여 비음수 행렬분해(nonnegative matrix factorization, NMF) 기반으로 다중음원의 도래각을 추정하는 새로운 방법을 제안한다. 우선 이중 마이크로폰 배열에 들어온 음향 신호들을 연속된 분석프레임으로 분할한 후, 각 프레임에 대해 조향응답파워 위상변환(steered-response power phase transform, SRP-PHAT) 빔형성기를 적용하여 스테레오 신호들을 시간-방향 영역으로 표현한다. 이러한 SRP-PHAT의 시간-방향 출력값들은 사전에 정의된 프레임 수만큼 누적하여 시간-방향 블록으로 정의한다. 다음으로, 잡음에 강건한 도래각 추정을 위하여, 각 시간-방향 블록을 블록차감 기법을 사용하여 매 프레임에 대해 정규화한다. 이후, 다중음원 환경에서 각 음원의 방향을 클러스터링하기 위해 정규화된 시간-방향 블록에 비지도(unsupervised) NMF를 적용한다. 구체적으로, 음원의 개수와 이들의 도래각을 추정하는데 각각 활성 및 기저 행렬들을 사용한다. 제안된 방법의 도래각 추정 성능을 평가하기 위해 이중 마이크로폰 배열로부터 입력된 [-35◦, 5m], [12◦, 4m], 그리고 [38◦, 4.m]에 각각 위치한 세 가지 음원들에 대한 추정 오차의 절대 평균(mean absolute error, MAE) 및 오차의 표준편차를 측정하였다. 실험 결과. 제안된 방법은 기존의 SRP-PHAT 기반 도래각 추정방법에 비해 상대적으로 MAE를 56.83% 줄일 수 있었다.

Abstract

This paper proposes a new nonnegative matrix factorization (NMF) based direction-of-arrival (DOA) estimation method for multiple sound sources using a dual microphone array. First of all, sound signals coming from the dual microphone array are segmented into consecutive analysis frames, and a steered-response power phase transform (SRP-PHAT) beamformer is applied to each frame so that stereo signals of each frame are represented in a time-direction domain. The time-direction outputs of SRP-PHAT are stored for a pre-defined number of frames, which is referred to as a time-direction block. Next, In order to estimate DOAs robust to noise, each time-direction block is normalized along the time by using a block subtraction technique. After that, an unsupervised NMF method is applied to the normalized time-direction block in order to cluster the directions of each sound source in a multiple sound source environments. In particular, the activation and basis matrices are used to estimate the number of sound sources and their DOAs, respectively. The DOA estimation performance of the proposed method is evaluated by measuring a mean absolute error (MAE) and the standard deviation of errors between the oracle and estimated DOAs under a three source condition, where the sources are located in [-35◦, 5m], [12◦, 4m], and [38◦, 4.m] from the dual microphone array. It is shown from the experiment that the proposed method could relatively reduce MAE by 56.83%, compared to a conventional SRP-PHAT based DOA estimation method.

Keywords : Multiple DOAs, GCC-PHAT, SRP-PHAT, NMF

* 학생회원, ** 평생회원, 광주과학기술원 전기전자컴퓨터공학부 (School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology)
*** 정회원, 한국통신 융합기술원 서비스 연구소 (Service Laboratory, Institute of Convergence Technology, Korea Telecom)
© Corresponding Author (E-mail : hongkook@gist.ac.kr)
※ 이 논문은 국토부의 재원으로 국토교통과학기술진흥원의 지원을 받아 수행된 연구사업임 (16TBIP-C111209-01)
Received ; October 12, 2016 Revised ; January 9, 2017 Accepted ; January 19, 2017

Ⅰ. 서 론

센서 배열을 이용한 도래각 추정(direction-of-arrival, DOA)은 안테나 어레이를 이용한 무선통신[1], 또는 마이크로폰 배열을 이용한 음원추정 및 개선[2] 등에 응용될 수 있다. 특히 음원추정 및 개선을 위한 도래각 추정은 빔형성 기반의 전처리(pre-processing)로써 활용될 수 있으며, 이러한 기술은 최근의 Amazon의 Echo와 같은 원거리 음성인식 장치에 주요하게 사용된다.[3] 도래각 추정 기법의 또 다른 응용 분야로는 보안카메라의 사각 문제를 해결하는 데 활용될 수 있다.[4] 즉, 화각으로 인지할 수 없는 범위의 사건 현장을 사람의 비명 소리, 자동차의 충돌 및 충격음을 감지하고 방향을 추정한 후, 카메라를 추정된 방향으로 회전시켜 사각으로 인한 문제를 해결할 수 있게 해준다. 위와 같은 이점과 다양한 응용 분야를 갖는 도래각 추정 기술은 실생활에서 매우 필요한 기술임을 알 수 있다.[3~5]

도래각은 입력되는 다채널의 신호들 사이에 상호상관(cross correlation, CC)을 기반으로 가장 상관도가 높게 될 때의 lag가 도래시간차(time difference of arrival, TDOA)로 부터 추정될 수 있다.[2] 하지만 이 경우 잔향이나 잡음 등의 영향에 의해 DOA 추정 성능이 열화되는 경향이 있다.[6] 이러한 문제를 극복하기 위하여, 시간영역에서의 CC는 주파수 영역에서의 상호 파워스펙트럼으로 표시되며, 이를 통해 상호상관함수에서 시간지연을 추정하는 대신 상호상관함수를 적절한 사상함수를 이용하여 공간좌표로 사상시킨 후 음원의 위치를 추정하는 일반화된 상호상관(generalized CC, GCC) 방법이 널리 사용되고 있다.[6] GCC에서 사상함수에 적용할 수 있는 다양한 가중치가 존재하게 되는데, 그 중 위상변환(phase transform, PHAT)은 잔향 환경(reverberant condition)에서 강건한 성능을 보임이 확인되었다.[6] 즉, CC 기반의 DOA 방법에 비해 구현의 용이성과 정확도를 강점으로 갖는 GCC에 더해서 PHAT 가중치를 적용함으로써 잔향 환경에 강건한 성능을 보이게 된다. 하지만 이 경우에도 위상차의 파워가 큰 음원에 의존적이어서 다중음원에 대한 방향추정 시 그 성능이 떨어지게 된다.[10]

기존의 방법을 보완한 방법으로 조향응답파워(steered response power, SRP) 기반의 방향추정 방법이 있다.[7] 특히 SRP에 PHAT 가중치를 적용한 SRP-PHAT의 강점은 잔향 환경에서 강건한 성능을 보이며 GCC-PHAT보다 정확하다고 알려져 있다. 하지만 마이크로폰이 2개인 경우, SRP-PHAT은 GCC-PHAT과 등가가 되어 GCC-PHAT과 마찬가지로 다중음원의 DOA 추정 시 정확도가 떨어진다는 단점이 있다.[7~8]

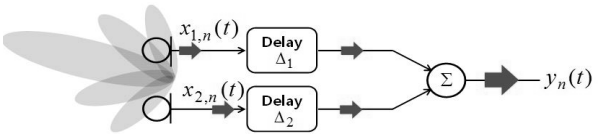

그림 1. 이중 마이크로폰을 이용한 조향응답파워 기반 도래방향추정

Fig. 1. Steered response power based DOA estimation using a dual microphone array.

따라서 본 논문에서는 한 쌍의 마이크로폰으로부터 존재하는 다수 음원의 위치를 추정하는 기법을 제안한다. 먼저, 제안된 기법은 배경잡음의 영향으로 발생할 수 있는 SRP-PHAT 영역의 허상을 감쇄하기 위해 블록차감법(block subtraction)을 적용한다. 다음으로, 다중음원이 생성하는 다수의 SRP-PHAT 피크 값들을 강건하게 분류하기 위해 SRP-PHAT 영역에 대한 비지도 학습(unsupervised learning) 기반의 비음수 행렬 분해(nonnegative matrix factorization, NMF)을 적용함으로써 다중음원의 도래각을 추정한다. SRP-PHAT과 NMF를 이용하여 음원의 위치를 추정하는 기존의 연구[11]이 있지만, 해당 기술은 위치를 추정하고자 하는 음원의 수를 사전에 알고 있다는 가정 하에 주어진 모든 프레임에 대한 off-line NMF를 수행하여 다중음원의 위치를 추정한다.[11] 이에 반해, 제안된 방법은 프레임 단위로 on-line NMF를 진행하여 매 프레임마다 음원의 수를 추정하고 이에 대해 각도 추정 결과를 도출하기 때문에, 음원의 수를 미리 알 수 없는 환경 및 음원의 수가 가변인 환경에 보다 적합하다. 또한 프레임 단위로 각도 추정 결과를 얻을 수 있어 실시간 응용에 용이하게 적용이 가능하다는 장점이 있다.

본 논문의 구성은 다음과 같다. Ⅱ절에서는 SPR-PHAT 기반의 음원 방향추정 기법을 기술한다. 그리고 Ⅲ절에서는 SRP-PHAT의 블록차감법과 NMF에 기반한 다중음원 방향추정 기법을 제안한다. Ⅳ절에서는 제안된 방법의 성능을 평가한 후, Ⅴ절에서는 본 논문의 결론을 맺는다.

Ⅱ. 기존의 음원 방향추정 방법

1. 원거리 신호모델

원거리에서 입사하는 신호모델은 한 점에서 발생해

서 입사하였으며 각각의 마이크에 대해서 에너지 손실은 없고 참조 마이크를 기준으로 거리에 따른 시간 지연만 존재한다고 가정한다. 이때, 마이크로 녹음되는 신호는 다음과 같이 표현된다.[6~7]

$$x_{m,n}(t) = s_n(t+\tau_{1,m}) * h_m(t) + v_n(t) \quad (1)$$

여기서, $x_{m,n}(t)$는 시간영역에서의 m번째 채널로 녹음되는 n번째 프레임의 신호를 의미한다. 시간지연, $\tau_{1,m}$은 마이크 간격, d와 신호의 입사각도, θ_{DOA}에 의해서 결정되며, $\tau_{1,m} = \frac{1}{c} d sin(\theta_{DOA})$로 표현될 수 있다. 이때 c는 공기 중에서의 음속을 의미한다. 그리고 $s_n(t)$와 는 시간 영역에서 깨끗한 신호와 배경잡음을 각각 $v_n(t)$의미한다. 그리고 $h_m(t)$는 m번째 채널의 실내 공간의 전달함수이다.[6~7]

2. SRP-PHAT 기반 방향추정 방법

그림 1은 SRP-PHAT 방법의 블록도이다. 그림에서 보는 바와 같이, SRP-PHAT은 일반적으로 지연합(delay and sum, DS) 빔형성 기법을 이용하여 임의의 방향, p에 해당하는 조향응답파워를 분석하여 음원의 발생 방향을 추정한다. 즉, p에 대해서 조향한 신호의 파워, $R_{p,n}$는 다음과 같이 정의될 수 있다.[7~8]

$$R_{p,n} = \sum_{k=1}^{M} \sum_{l=k+1}^{M} \int_{-\infty}^{\infty} \left[\Psi_{kl,n}(\omega) X_{k,n}(\omega) X_{l,n}^{*}(\omega) e^{j\omega(\tau_{k,l}(p))} \right] d\omega \quad (2)$$

여기서, M은 전체 채널수를 의미하고, $X_{m,n}(\omega)$은 m번째 채널에 녹음된 신호에 이산 푸리에 변환을 적용하여 얻은 n번째 프레임의 복소 스펙트럼을 의미한다. 또한 $\Psi_{k,l,n}(\omega) = 1/|X_{k,n}(\omega)X_{l,n}^{*}(\omega)|$는 PHAT 가중치를, $\tau_{k,l}(p)$은 p로부터 k번째 및 l번째 마이크로폰으로의 도달 시간 차이를 각각 의미한다. 즉, 음원의 위치는 $R_{p,n}$을 최대화하는 한 지점을, p^{*}라 할 때, 이는 다음과 같이 표현될 수 있다.[9]

$$p^{*} = \mathrm{argmax}_{p} R_{p,n}. \quad (3)$$

최종적으로, p^{*}로부터 음원의 위치에 해당하는 도래각, $\theta_{DOA} = \sin^{-1}(\tau_{1,m}(p^{*})c/d)$를 추정할 수 있다.

상기의 SRP-PHAT 방법은 배경잡음 및 잔향 환경에서 단일 음원을 찾는데 강건한 성능을 보이는 것으로 알려져 있다.[8~9] 하지만 두 개 이상의 동시에 발생하는 다중음원의 위치를 추정하는데 있어서 단순히 SRP-PHAT의 최대치를 찾는 방법은 다중음원이 발생시키는 true-peak과 허상 및 잡음이 발생시키는 false-peak 간의 구분에 취약하다는 단점을 지닌다.[10] 이러한 한계를 극복하기 위해, 다음 절에서는 NMF를 이용하여 다중음원의 SRP-PHAT 최대치들을 분별하는 방법을 제안하도록 한다.

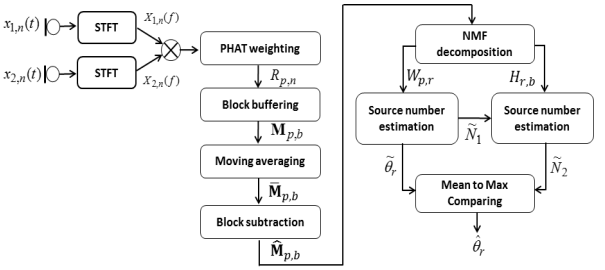

그림 2. 제안된 NMF 기반 다중음원 방향추정 블록도
Fig. 2. Block diagram of the proposed NMF-based DOA estimation of multiple sound sources.

Ⅲ. 제안된 NMF 기반 다중음원 방향추정 방법

본 절에서는 NMF 기반 다중음원 방향추정 방법을 제안한다. 그림 2는 제안된 방법의 블록도를 보여준다. 그림에서 보는 바와 같이, 제안된 방법은 두 개의 마이크로폰 배열로부터 SRP-PHAT을 추정한 다음, 허상에 의해 나타나는 SRP-PHAT의 false-peak 값들을 제거하기 위한 블록차감법을 수행한다. 허상이 제거된 SRP-PHAT은 NMF를 통해 기저 및 활성행렬로 분해되고, 이들 행렬들로부터 음원의 수를 추정하고 추정된 음원의 수만큼의 음원의 방향을 각각 추정한다.

1. 블록차감법

실제 환경에서 녹음된 음원의 SRP-PHAT는 배경잡음 등의 영향으로부터 허상에 의한 false-peak이 발생하게 된다.[11] 단일음원의 방향추정의 경우 이러한 false-peak과 true-peak 간의 크기 차이가 대체로 크게 나타나기 때문에 큰 문제가 안 되지만, 다중음원의 경우 false-peak 및 true-peak 간의 차이가 작아 방향추정 성능이 저하된다. 이를 해결하기 위해, 제안된 방법은 SRP-PHAT 영역에서의 블록차감법을 적용하여 일정 시간 동안 평균화된 SRP-PHAT의 크기보다 낮은 크기를 지니는 peak은 false-peak이라 간주하여 제거함으로써 상대적으로 강건한 peak들만 유지되도록 한다.

먼저, 주어진 SRP-PHAT를 시간-방향 블록단위로 처리하기 위해 B 길이의 버퍼에 SRP-PHAT 열벡터를 저장한다. 즉, $P\times B$ 차원의 시간-방향 블록은 다음과 같이 표현된다.

$$\mathbf{R}_n = [r_{p,n-B+1},\cdots,r_{p,n-b},\cdots,r_{p,n}] \quad (4)$$

여기서, p는 $\theta_{DoA}\in[-90^\circ, 90^\circ]$를 만족하는 탐색 범위를 지니도록 설정된다. 즉, $p=\tau_{1,m}^{-1}((d/c)\sin(\theta_{DOA}))$로 표현되며 그 최대치는 P이다. 다음으로, 시간-방향 블록에 대한 시간-방향 평균을 구하여 n번째 SRP-PHAT 프레임에 대한 차감량을 다음 식과 같이 계산한다.

$$\bar{r}_n = \frac{1}{PB}\sum_{p=1}^{P}\sum_{b=1}^{B} r_{p,n-b+1}. \quad (5)$$

그러고 나서, $\mathbf{R}_n$에 대해 다음 식과 같이 차감된다.

$$\hat{r}_{p,n-b+1} = \begin{cases} r_{p,n-b+1}-\bar{r}_n & \text{if } r_{p,n-b+1} > \bar{r}_n \\ \beta & \text{otherwise} \end{cases} \quad (6)$$

여기서, b의 범위는 0부터 B-1까지이며, β는 0보다 큰 바닥계수로써 본 논문에서는 0.00001로 설정하였다. 최종적으로 블록차감법이 적용된 시간-방향 블록, $\hat{\mathbf{R}}_{p,n} = [\hat{r}_{p,n-B},\cdots,\hat{r}_{p,n-b},\cdots,\hat{r}_{p,n}]$이 얻어지고, 이는 NMF 기반 다중음원 방향추정에 활용된다.

2. NMF 기반 다중음원 방향추정

NMF[12~14]는 다변수의 자료 구조를 분해하는데 있어 효과적인 수단으로써, 본 논문에서는 다수의 SRP-PHAT peak 값을 분리하여 다중음원의 DOA 추정 정확성을 높이는 데 활용된다. 특히, 제안된 방법은 비지도 NMF 기법을 SRP-PHAT 블록에 적용하여 음원의 개수와 이에 대한 각각의 음원 위치를 추정한다. 구체적으로, 먼저 SRP-PHAT 블록에 대하여 다음과 같이 비음수 행렬 분해를 수행한다.

$$\hat{\mathbf{R}}_{p,n} \simeq \mathbf{W}_{p,r}\mathbf{H}_{r,b} \quad (7)$$

여기서, $\mathbf{W}_{p,r}$는 NMF에 의해 분해된 $P\times R$ 차원의 기저행렬(basis matrix)이며 r차원 번째 음원에 따른 각도, p와의 관계를 의미한다. 또한 $R\times B$ 차원 행렬, $\mathbf{H}_{r,b}$를 활성행렬(activation matrix)라 하고 $\mathbf{H}_{r,b}$의 b번째 열은 $\hat{\mathbf{R}}_{p,n}$의 b번째 열벡터에서 가장 활동적인 r번째 음원을 나타낸다. 식 (7)의 행렬 $\mathbf{W}_{p,r}$와 $\mathbf{H}_{r,b}$는 Kullback-Leibler (KL) divergence를 최소화하는 방향으로 아래의 식 (8)과 (9)의 반복 연산을 통해 업데이트 된다.[14]

$$\mathbf{W}_{p,r}^{(i)} \leftarrow \mathbf{W}_{p,r}^{(i-1)} \otimes \frac{(\boldsymbol{\Lambda}_{p,b}^{-1^{(i-1)}}\otimes\hat{\mathbf{R}}_{p,n})\mathbf{H}_{r,b}^{T^{(i-1)}}}{\mathbf{H}_{r,b}^{T^{(i-1)}}} \quad (8)$$

$$\mathbf{H}_{r,b}^{(i)} \leftarrow \mathbf{H}_{r,b}^{(i-1)} \otimes \frac{\mathbf{W}_{p,r}^{T^{(i)}}(\hat{\mathbf{R}}_{p,n}\otimes\boldsymbol{\Lambda}_{p,b}^{-1^{(i-1)}})}{\mathbf{W}_{p,r}^{T^{(i)}}+\mu} \quad (9)$$

여기서, $\otimes$, /, 그리고 T는 각각 행렬의 원소 곱, 원소 나눗셈, 그리고 전치행렬 연산을 각각 의미하며, i는 식 (8)과 (9)의 반복 인덱스를 나타낸다. 또한 $\boldsymbol{\Lambda}_{p,b}^{-1^{(i)}} = 1/\mathbf{W}_{p,r}^{(i)}\mathbf{H}_{r,b}^{(i)}$을 의미한다.

식 (8)과 (9)의 반복 연산에 앞서, $\mathbf{H}_{r,b}^{(1)}$의 모든 요소 값들은 0에서 1사이의 균등 분포를 따르는 난수로 초기화된다. 또한, 기저벡터에 한 음원의 방향을 나타내기 위해 기저행렬의 초기값, $\mathbf{W}_{p,r}^{(1)}$을 다음과 같이 정의한다.

$$\mathbf{W}_{p,r}^{(1)} = [\mathbf{g}^1(\hat{p}(1))^T,\cdots,\mathbf{g}^r(\hat{p}(r))^T,\cdots,\mathbf{g}^R(\hat{p}(R))^T] \quad (10)$$

여기서, $\boldsymbol{g}^r(\hat{p}(r))$는 $\hat{p}(r)$에서 최대값, 즉 1을 지니는 P 길이의 Hamming window로써 $\mathbf{W}_{p,r}^{(1)}$의 r번째 기저벡터에 해당한다. 그리고 $\hat{p}(r)$은 $\hat{\boldsymbol{R}}_{p,n}$의 상위 R개의 peak에 해당하는 p를 의미한다. 식 (8)과 (9)에 표현된 NMF 연산은 KL divergence의 목적함수 값이 사전에 정의된 문턱값보다 낮아질 때까지 반복된다. 이 반복 연산의 종료 시점, 즉 $i=I$에서 $\mathbf{W}_{p,r}=\mathbf{W}_{p,r}^{(i=I)}$와 $\mathbf{H}_{r,b}=\mathbf{H}_{r,b}^{(i=I)}$를 얻는다.

다음으로, 제안된 방법은 분해된 기저행렬, $\mathbf{W}_{p,r}$의 열벡터 간 유사도를 비교하여 유효한 음원의 수를 추정한다. 이를 위해, $\mathbf{W}_{p,r}$을 구성하는 열벡터, $\mathbf{w}_{p,r}$과 유사한 $\mathbf{w}_{p,r'}$의 존재 여부에 대한 표식자(indicator), $D(r)$을 다음 식과 같이 구한다.

$$D(r) = \begin{cases} 1 \;\; if\; |\tilde{p}(r)-\tilde{p}(r')| > \epsilon_1 \\ 0 \;\; otherwise \end{cases} \quad (11)$$

여기서, $r'\in[1,R]$, $r\neq r'$이고, ϵ_1은 유사도 결정 파라메터이다. 그리고 식 (11)에서 $\tilde{p}(r)=\mathrm{argmax}_p(\mathbf{w}_{p,r})$로 정의된다. 또한, 제안된 방법은 $\mathbf{H}_{r,b}$의 r번째 행의

크기, $h(r)$의 평균과 최대값의 비율(mean-to-max ratio)을 사용하여 다중음원 방향추정에 사용하는 r^*를 얻는다. 즉, 모든 $r \in [1, R]$에 대해서 다음 식을 만족하는 r^*을 찾는다.

$$h(r)/(\frac{1}{R}\sum_{r=1}^{R} h(r)) > \epsilon_2 \; and \; D(r) = 1 \tag{12}$$

여기서, ϵ_2는 mean-to-max ratio에 대한 threshold를 나타낸다. 식 (12)를 만족하는 r^*의 개수로 다중음원의 수를 추정한다. 마지막으로, 다중음원의 위치 추정은 r^*의 각각에 대해 다음과 같이 얻는다.

$$\theta_{DOA}(r^*) = \sin^{-1}(\tau_{1,m}(\tilde{p}(r^*))c/d). \tag{13}$$

Ⅳ. 성능 평가

제안된 방법의 성능을 평가하기 위해서 다음과 같이 실험 환경을 구성하였다. 즉, 음원은 IEEE Single- and Multichannel Audio Recordings Database (SMARD)를 사용하였다.[16] SMARD는 약 7시간 분량으로 구성되어 있으며 각각의 음원은 48 kHz로 표본화되고 각 샘플은 16bit 해상도를 갖는다. 녹음 환경은 약 0.15초의 잔향시간을 지니는 60m^{2}의 직사각형 형태의 실내 공간에 세 개의 라우드스피커가 고정 배치되어 음원이 재생되고 7 채널 선형 마이크로폰 배열로 녹음되었다.

본 논문에서는 이중에서 5cm 간격으로 떨어진 3, 4번 두 개의 채널을 선택하여 사용하였다. 또한, 음원의 위치는 마이크배열 전방 기준으로 각각 (A) [-35◦, 5m], (B) [12◦, 4m], 그리고 (C) [38◦, 4.m]이였으며, 각 위치별로 음성, 보컬, 그리고 음악 콘텐츠 각각 6가지 음원 clip을 모두 합쳐 총 7분 분량을 재생하였다. 음원수에 따른 성능 평가 시, 단일음원의 경우 (A), (B), 그리고 (C)를, 이중음원의 경우 (A, B), (A, C), 그리고 (B, C)를, 그리고 삼중음원의 경우 (A, B, C)의 음원을 사용하였다.

제안된 방법의 성능은 기존의 SRP-PHAT[8] 방법과 비교하였다. 특히, 제안된 방법의 요소 기술인 SRP-PHAT 블록차감법 및 NMF 기반 방법이 방향추정 성능에 미치는 영향을 분석하기 위해 SRP-PHAT에 각각의 세부 기술만 적용한 방법과도 그 성능을 비교하였다. 실험의 진행을 위한 제안된 방법의 파라메터들은 위에 기술한 평가용 음원과 별개로 준비된 개발용 샘플 음원들에 대한 최적값을 찾아 반영하였으며, 그 값은 각각 $B = 5$, $P = 181$, $R = 10$, $\epsilon_1 = 5$, 그리고 $\epsilon_2 = 2$였다. 성능 평가에 대한 척도로는 음원의 실제 DOA와 DOA 추정치의 차이간 오차의 절대 평균 (means absolute error, MAE) 및 이의 표준편차를 사용하였다.[18]

표 1. 음원수에 다른 DOA추정 오차의 평균 및 표준편차 (°)
Table1. Mean and standard deviation (°) of DOA estimation errors for different number of sound sources.

음원수	SRP	SRP+BT	SRP+ NMF	SPR+BT+ NMF (Proposed)
1	3.34 (2.2847)	3.02 (1.5937)	2.54 (1.1576)	2.48 (1.0583)
2	9.92 (4.0626)	5.41 (3.0008)	5.12 (3.5735)	3.28 (2.8275)
3	9.37 (4.7697)	8.91 (3.6715)	7.41 (3.8464)	4.05 (3.1321)

표 2. 음원수에 따른 기존 방법 대비 제안된 방법의 DOA 추정 오차의 상대적 감소율 (%)
Table2. Relative reduction of DOA estimation error (%) for different number of sound sources.

음원수	SRP+BT vs SRP	SRP+NMF vs SRP	Proposed vs SRP
1	9.58	23.95	25.75
2	45.44	48.41	66.92
3	4.91	20.92	56.83

$$e_{DOA} = \frac{1}{N}\sum_{n=1}^{N} |\theta_{ref}(n) - \theta_{esti}(n)| \tag{14}$$

$$\sigma_{DOA} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} (|\theta_{ref}(n) - \theta_{esti}(n)| - e_{DOA})^2} \tag{15}$$

본 논문의 실험에서 DOA 추정은 한 개 이상의 음원이 존재하는 구간에서만 이루어졌다. 따라서 식 (14) 및 (15)의 N은 평가에 사용된 음원 존재 구간에 대한 총 프레임의 수를 의미하며, $\theta_{ref}(n)$와 $\theta_{esti}(n)$는 음원의 n번째 프레임의 실제 DOA와 추정된 DOA를 각각 나타내며, 그 단위는 degree (°)이다. 본 논문에서 약 2만 프레임(N=20000)에 대한 MAE를 측정하였다.

표 1은 음원수에 따른 각 방법의 음원 방향추정 오차의 평균과 괄호안의 숫자로 표기된 표준편차를 보여준다. 또한, 표 2는 기존 방법들 대비 제안된 방법의 상대적 위치추정 오차 감소율을 보여준다. 표 1 및 2에서 보는

바와 같이, 제안된 NMF 기반 다중음원 방향추정 방법은 기존의 SRP-PHAT 방법들에 비해 낮은 추정 오차를 보였다. 특히, 제안된 방법은 SRP-PHAT 대비 이중 및 삼중음원 조건에서 각각 66.92%와 56.83%의 높은 상대적 오차 감소율을 보였다. 또한, 제안된 방법의 두 가지 요소기술인 SRP-PHAT 블록차감법 (SRP+BS)과 NMF (SRP+NMF) 모두 기존의 SRP에 비해 추정 오차를 감소시키는 것을 볼 수 있다. 특히, SRP+NMF가 SRP+BS에 비해 상대적으로 효과적임을 알 수 있다.

Ⅳ. 결 론

본 논문에서는 비음수 행렬 기반의 다중음원 방향추정 방법을 제안하였다. 제안된 방법은 강건한 다중음원 위치 추정을 위해 SRP-PHAT의 시간-방향 영역에 블록차감법 및 NMF 기법을 적용하였다. 제안한 방법의 성능을 검증하기 위해서 SMARD 데이터베이스를 활용하여 음원의 각도와 개수에 따른 오차의 절대 평균 및 표준편차를 측정하였다. 실험 결과, 제안된 방법이 다중음원 조건에서 종래의 SRP-PHAT 방법 대비 MAE를 56.83% 줄일 수 있었고, 음원의 수가 증가할수록 NMF가 효과적임을 알 수 있었다.

REFERENCES

[1] M. J. Kim, “Direction of arrival estimation in colored noise using wavelet decomposition,” Journal of The Institute of Electronics and Information Engineers, Vol. 37, No. 11, pp. 48-59, Nov. 2000.

[2] M. Brandstein and D. Ward, Microphone Arrays: Signal Processing Techniques and Applications, Springer Science & Business Media, 2001.

[3] M. Vacher, B. Lecouteux, J. S. Romreo, M. Ajili, F. Portet, and S. Rossato, “Speech and speaker recognition for home automation: Preliminary results,” in Proc. of International Conference on Speech Technology and Human-Computer Dialogue (SPeD), Bucharest, Romania, pp. 181-190, Oct. 2015.

[4] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, “Scream and gunshot detection and localization for audio-surveillance systems,” in Proc. of IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS), London, UK, pp. 21-26, Sept. 2007.

[5] K. Ishiguro, T. Yamada, S. Araki, and T. Nakatani, “A probabilistic speaker clustering for DOA-based diarization,” in Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, pp. 241-244, Oct. 2009.

[6] C. H. Knapp and G. C. Carter, “The generalized correlation method for estimation of time delay,” IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 24, No. 4, pp. 320-327, Aug. 1976.

[7] D. H. Johnson and D. E. Dudgeon, Array Signal Processing: Concepts and Techniques: Simon & Schuster, 1992.

[8] H. Do, H. F. Silverman, and Y. Yu, “A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array,” in Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Honolulu, HI, pp. 121-124, Apr. 2007.

[9] J. P. Dmochowski, J. Benesty, and S. Affes, “A generalized steered response power method for computationally viable source localization,” IEEE Transactions on Audio, Speech, and Language Processing, Vol. 15, No. 8, pp. 2510-2526, Nov. 2007.

[10] H. Kayser, J. Anemuller, and K. Adiloglu, “Estimation of inter-channel phase differences using non-negative matrix factorization,” in Proc. of IEEE 8th Sensor Array and Multi-channel Signal Processing Workshop (SAM), A Coruña, Spain, pp. 77-80, June 2014.

[11] J. Traa, P. Smaragdis, N. D. Stein, and D. Wingate, “Directional NMF for joint source localization and separation,” in Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, pp. 1-5, Oct. 2015.

[12] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in Proc. of Neural Information Processing System (NIPS), Denver, CO, pp. 556-562, Dec. 2000.

[13] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” Nature, Vol. 401, No. 6755, pp. 788-791, Nov. 1999.

[14] J. Le Roux, F. J. Weninger, and J. R. Hershey, Sparse NMF - half-baked or Well Done?, Mitsubishi Electric Research Labs (MERL), Cambridge, MA, Technical Report TR-2015-23, Mar. 2015.

[15] A. H. Nuttall, “Some windows with very good sidelobe behavior,” IEEE Transactions on Acoustics,

Speech and Signal Processing, Vol. 29, No. 1, pp. 84-91, Jan. 1981.
[16] J. K. Nielsen, J. R. Jensen, S. H. Jensen, and M. G. Christensen, "The single-and multichannel audio recordings database (SMARD)," in Proc. of 14th International Workshop on Acoustic Signal Enhancement (IWAENC), Antibes, France, pp. 40-44, Sept. 2014.
[17] P. Aarabi and G. Shi, "Phase-based dual-microphone robust speech enhancement," IEEE Transactions on Systems, Man, and Cybernetics -Part B: Cybernetics, Vol. 34, No. 4, pp. 1763-1773, Aug. 2004.
[18] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," International Journal of Forecasting, Vol. 22, No. 4, pp. 679-688, May 2006.

저 자 소 개

전 광 명(학생회원)
2010년 2월 세종대학교 정보통신공학과 학사
2012년 2월 광주과학기술원 정보통신공학부 석사
2012년 3월～현재 광주과학기술원 전기전자컴퓨터공학부 박사과정
<주관심분야: 음성 및 오디오 신호처리, 기계학습, 딥러닝>

유 승 우(정회원)
2014년 2월 동국대학교 IT학부 전자공학과 학사
2016년 2월 광주과학기술원 정보통신공학부 석사
2016년 8월～현재 KT 융합기술원 서비스 연구소 연구원
<주관심분야: 음성 및 오디오 신호처리, 음성인식>

김 홍 국(평생회원)
1988년 2월 서울대학교 제어계측공학과 학사
1990년 2월 한국과학기술원 전기 및 전자공학과 석사
1994년 8월 한국과학기술원 전기 및 전자공학과 박사
1990년～1998년 삼성종합기술원 전문연구원
1998년～1998년 MMC Technology 선임연구원
1998년～2003년 AT&T Labs-Research Senior Member Technical Staff
2014년～2015년 City University of New York, Visiting Professor
2003년 8월～현재 광주과학기술원 전기전자컴퓨터공학부 교수
<주관심분야: 음성인식, 음성 및 오디오 신호처리, 3D 오디오, 딥러닝>