2013년 12월 전자공학회 논문지 제 50 권 제 12 호
Journal of The Institute of Electronics Engineers of Korea Vol. 50, NO. 12, December 2013 http://dx.doi.org/10.5573/ieek.2013.50.12.191

논문 2013-50-12-22

인접 배치된 스테레오 무지향성 마이크로폰 환경에서 양이간 강도차를 활용한 음원 분리 기법

(Sound Source Separation Using Interaural Intensity Difference in Closely Spaced Stereo Omnidirectional Microphones)

전 찬 준*, 정 석 희*, 김 홍 국**

(Chan Jun Chun, Seok Hee Jeong, and Hong Kook Kim©)

요 약

본 논문에서는 실제 환경에서 인접 배치된 무지향성 스테레오 마이크로폰을 활용하여 녹음받은 스테레오 오디오 신호를 양이간 강도차에 기반하여 원하는 방위각에 존재하는 음원을 추출하는 음원 분리 기법을 제안한다. 먼저, 최소 분산 무손실 응답 빔형성기를 활용하여 스테레오 오디오 신호의 양이간 강도차를 극대화하고, 강도차 기반의 음원 분리 기법을 적용한다. 제안된 기법의 성능을 검증하기 위하여 stereo audio source separation evaluation campaign (SASSEC)에서 제공하는 객관적 성능평가 지표인 source-to-distortion ratio (SDR), source-to-interference ratio (SIR), sources-to-artifacts ratio (SAR)을 측정하였다. 측정한 결과, 음원 분리 기법에 빔형성기까지 적용한 결과가 높은 성능을 보인 것으로 평가되었다.

Abstract

In this paper, the interaural intensity difference (IID)-based sounr source separation method in closely spaced stereo omnidirectional microphones is proposed. First, in order to improve the channel separability, a minimum variance distortionless response (MVDR) beamformer is employed to increase the intensity difference between stereo channels. After that, IID-based sound source separation method is applied. In order to evaluate the performance of the proposed method, source-to-distortion ratio (SDR), source-to-interference ratio (SIR), and sources-to-artifacts ratio (SAR), which are defined as objective evaluation criteria in stereo audio source separation evaluation campaign (SASSEC), are measured. As a result, it was shown from the objective evaluation that the proposed method outperforms a sound source separation method without applying a beamformer.

Keywords : Sound source separation, minimum variance distortionless response (MVDR) beamformer, interaural intensity difference (IID)

* 학생회원, ** 평생회원, 광주과학기술원 정보통신공학부
(Gwangju Institute of Science and Technology, School of Information and Communications)
© Corresponding Author(E-mail: hongkook@gist.ac.kr)
※ 본 연구는 미래창조과학부 및 정보통신산업진흥원의 대학 IT연구센터 지원사업(NIPA-2013-H0301-13-4005) 및 2013년도 미래창조과학부의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2012-010636).
접수일자: 2013년10월16일, 수정완료일: 2013년11월24일

Ⅰ. 서 론

음원 분리 기술은 오디오 분야에서 많이 연구가 진행되고 있는 토픽 중에 하나이다. 지금까지 여러 알고리즘이 연구되었고, 지금도 연구가 활발히 진행 중이다[1~4]. 그 중에서도 독립 성분 분석(independent component analysis, ICA) 알고리즘은 오디오 신호들간의 상호독

립적이며, non-Gaussian 이라는 가정을 통하여 음원을 분리한다[1]. Computational auditory scene analysis (CASA) 알고리즘에서는 인체 청각 시스템의 메커니즘을 기반으로 음원 분리 기술이 이루어진다[2~3]. 또한, azimuth discrimination and resynthesis (ADRess) 알고리즘에서는 스테레오 오디오 신호의 강도차를 활용하여 음원 분리 기술이 이루어진다[4]. 특히, ADRess 알고리즘은 스테레오로 레코딩된 음원에서 음원의 개수에 제한없이 채널간의 강도차를 활용하여 음원을 분리하며, 실시간으로 알고리즘이 구현 가능한 장점을 가진다. 하지만, ADRess 알고리즘은 스테레오 오디오 신호의 강도차가 분명하게 나타나고, 시간차나 위상차가 존재하지 않을 때 음원 분리가 가능한 제약이 있다. 즉, 실제 환경에서 인접 배치된 무지향성 스테레오 마이크로폰을 활용하여 녹음받은 음원의 경우에는 채널간의 강도차가 패닝된 음원에 비하여 분명하게 나타나지 않으며 시간차와 위상차 존재하게 되는데, 이렇게 실제 환경에서 녹음받은 음원의 경우에는 ADRess 알고리즘이 패닝된 음원에 비하여 상대적으로 낮은 음원 분리 성능을 보인다.

본 논문에서는 ADRess 알고리즘이 실제 환경에서 인접 배치된 무지향성 스테레오 마이크로폰을 활용하여 녹음 받은 음원에서 강도차를 활용하여 음원 분리하는 방법을 제안한다. 구체적으로는 ADRess 알고리즘 이전에 최소 분산 무손실 응답 (minimum variance distortionless response, MVDR) 빔형성기 적용을 통하여 실제 환경에서 녹음받은 음원에 대해서 음원 분리 성능을 향상시킨다[5]. 즉, 빔형성기를 활용하여 스테레오 무지향성 마이크로폰으로 녹음된 오디오 신호의 강도차를 극대화시키고, 빔형성기를 통과한 신호에 대해 ADRess 알고리즘을 적용한다.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 무지향성 스테레오 마이크로폰이 인접 배치된 환경에서 강도차를 활용하여 음원 분리하는 기법을 제안한다. 그리고 Ⅲ장에서는 제안하는 음원 분리 기법의 성능을 검증하고, Ⅳ장에서 본 논문의 결론을 맺는다.

Ⅱ. 양이간 강도차를 활용한 음원 분리 기법

본 논문에서 제안하는 음원 분리 기법의 전체 구성도는 <그림 1>과 같다. <그림 1>에서 $x_L(n)$과 $x_R(n)$은

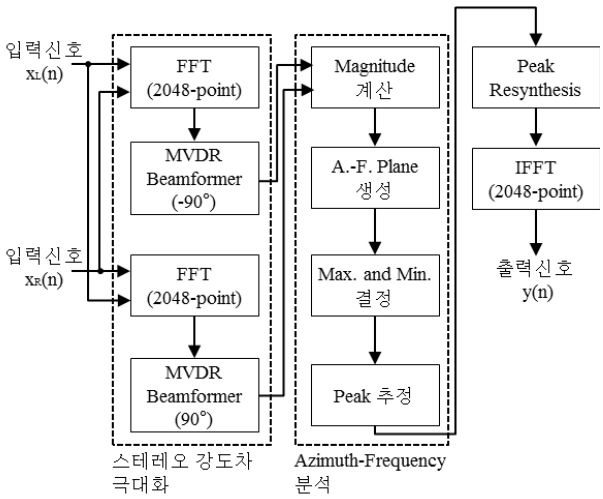


그림 1. 제안된 음원 분리 기법의 블록도
Fig. 1. Block diagram of the proposed sound source separation method.

인접 배치된 무지향성 스테레오 마이크로폰의 왼쪽과 오른쪽 채널에 해당하는 입력 신호이다. 먼저, 획득한 입력 신호를 MVDR 빔형성기를 적용하여 스테레오 오디오 신호에 대한 채널간 강도차를 극대화시킨다[5]. 그 다음으로 빔형성기를 통과한 스테레오 오디오 신호를 강도차에 따른 azimuth-frequency 분석을 하여 강도차에 따라서 음원의 방위각을 분석하고, 원하는 방위각에 대한 정보만을 취득하여 음원을 분리한다. 각 부분에 대한 자세한 설명은 다음과 같다.

1. MVDR 빔형성기를 활용한 스테레오 오디오 신호의 강도차 극대화

인접 배치된 무지향성 마이크로폰으로 받은 입력 신호는 아래의 수식과 같이 표현 가능하다.

$$\begin{bmatrix} x_L(n) \\ x_R(n) \end{bmatrix} = \begin{bmatrix} a_L s(n-\tau_L) \\ a_R s(n-\tau_R) \end{bmatrix} + \begin{bmatrix} v_L(n) \\ v_R(n) \end{bmatrix} \quad (1)$$

여기서 $s(n)$은 입력 신호로부터 분리하고자 하는 타겟 신호이며, $v_L(n)$과 $v_R(n)$은 왼쪽 채널과 오른쪽 채널에 해당하는 노이즈이며, $x_L(n)$과 $x_R(n)$이 스테레오 마이크로폰을 통하여 입력되는 오디오 신호이다. 또한, 여기서 a_L과 a_R은 타겟 신호가 마이크로폰으로 입력되면서 감쇄 정도를 나타내는 인자이며, τ_L과 τ_R은 지연 정도를 나타내는 인자이다. 수식 (1)을 N-point fast Fourier transform (FFT)를 통하여 주파수 도메인으로 표현하면 아래의 수식과 같다.

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

여기서 $\boldsymbol{X}^T=[X_L(k)\ X_R(k)]$과 $\boldsymbol{V}^T=[V_L(k)\ V_R(k)]$를 의미하며, $S(k)$는 $s(n)$의 k번째 주파수 성분을 나타낸다. 또한, $\boldsymbol{d}$는 $s(n)$의 스테레오 마이크로폰으로 입력받을 때, 방위각에 따라 나타나게 되는 감쇄와 지연 정도를 표현하는 벡터를 가리킨다.

$$\boldsymbol{d}^T=\left[a_L\exp(-j\frac{2\pi k\tau_L}{N})\ \ a_R\exp(-j\frac{2\pi k\tau_R}{N})\right] \tag{3}$$

본 논문에서 무지향성 스테레오 마이크로폰이 인접 배치된 환경인 것을 고려할 때, 두 마이크로폰의 감쇄 정도는 거의 동일하게 나타나며, 방위각에 따라 나타나는 지연 정도는 상대적으로 표현 가능한다. 즉, 수식 (3)을 아래의 수식처럼 간략화가 가능하다.

$$\boldsymbol{d}^T=\left[1\ a_R\exp(-j\frac{2\pi f_s k}{N}\frac{l}{c}sin\theta)\right] \tag{4}$$

여기서 f_s는 샘플링 주파수를 나타나며, l은 스테레오 마이크로폰의 간격을, c와 θ는 각각 소리의 속도와 음원의 방향을 나타낸다. 본 논문에서는 샘플링 주파수를 48 kHz, 마이크로폰 간격을 5 cm, 소리의 속도를 340 m/s로 설정하였다. 또한, 입력되는 오디오 신호를 주파수 도메인으로 fast Fourier transform하기 위하여 2048 샘플씩 한 프레임이 되도록 분할하였고, 현재 프레임이 이전 프레임과 1536 샘플씩 겹치도록 쉬프트하였다.

주파수 도메인에서 수행되는 일반적인 빔형성기는 아래의 수식처럼 입력 신호에 가중 벡터를 선형 조합하는 형태로 결정된다.

$$\hat{S}(k)=\boldsymbol{W}^{\boldsymbol{H}}\boldsymbol{X}=[\,W_1(k)\ \ W_2(k)\,]\begin{bmatrix}X_L(k)\\X_R(k)\end{bmatrix} \tag{5}$$

여기서 $\boldsymbol{W}$가 빔형성기의 가중 벡터를 나타내며, $^{\boldsymbol{H}}$는 Hermitian 연산자를 가리킨다. 본 논문에서는 스테레오 입력 신호의 강도차를 극대화하는 목적으로 빔형성기를 활용하였다. 즉, 왼쪽 채널과 오른쪽 채널의 강도차를 분명하게 나타나기 위하여 왼쪽과 오른쪽 방향에 해당하는 두 개의 빔형성기를 활용하였다.

$$\widehat{S_L}(k)=\boldsymbol{W_L^H X}\ \text{and}\ \widehat{S_R}(k)=\boldsymbol{W_R^H X} \tag{6}$$

여기서 $\boldsymbol{W_L}$과 $\boldsymbol{W_R}$은 왼쪽과 오른쪽 방향에 해당하는 가중 벡터를 나타낸다. MVDR 빔형성기는 원하는 방향에 대한 신호의 크기를 일정하게 유지하면서, 나머지 방향에 대한 신호의 크기를 최소화하는 형태로 빔을 형성한다[5].

$$\min \boldsymbol{W_m^H R_{XX} W_m}\ \text{subject to}\ \boldsymbol{W_m^H d_m}=1(m=L\text{or}R) \tag{7}$$

여기서 $\boldsymbol{R_{XX}}$는 입력된 스테레오 오디오 신호의 자기상관행렬을 나타내며, $\boldsymbol{d}_m$은 m번째 마이크로폰의 조향 벡터를 가리킨다. 본 논문에서는 왼쪽과 오른쪽 채널에 대한 조향 벡터의 방향을 -90도 및 90도로 각각 설정하였다. 즉, 수식 (3)에 -90도와 90도로 대입하여 계산된 조향 벡터는 아래의 수식과 같다.

$$\begin{aligned}\boldsymbol{d_L}&=\left[1\ \exp(j\frac{2\pi\cdot 48000\cdot 0.05}{2048\cdot 340}k)\right]\\ \boldsymbol{d_R}&=\left[1\ \exp(-j\frac{2\pi\cdot 48000\cdot 0.05}{2048\cdot 340}k)\right]\end{aligned} \tag{8}$$

수식 (7)의 조건을 만족하는 가중 벡터를 찾는 것이 MVDR 빔형성 기법의 핵심이며, 이는 Lagrange multiplier를 활용하여 아래의 수식처럼 가중 벡터를 결정 가능하다[6].

$$\boldsymbol{W}_m=\frac{\boldsymbol{R_{XX}^{-1}d}_m}{\boldsymbol{d}_m^{\boldsymbol{H}}\boldsymbol{R_{XX}^{-1}d}_m} \tag{9}$$

여기서 $\boldsymbol{R_{XX}}$는 입력 신호의 자기상관행렬이므로 노이즈의 power spectral density matrix로 표현이 가능하다[7]. 즉, 수식 (9)를 아래의 수식처럼 변경이 가능하다.

$$\boldsymbol{W_m}=\frac{\boldsymbol{\Gamma_{VV}^{-1}d_m}}{\boldsymbol{d_m^H\Gamma_{VV}^{-1}d_m}} \tag{10}$$

여기서 $\boldsymbol{\Gamma_{VV}}$는 노이즈의 power spectral dinsity matrix로 아래의 수식처럼 계산된다.

$$\boldsymbol{\Gamma_{VV}}=\begin{bmatrix}1+\mu & sinc\left(2\pi f_s\frac{k}{N}\frac{l}{c}\right)\\ sinc\left(2\pi f_s\frac{k}{N}\frac{l}{c}\right) & 1+\mu\end{bmatrix} \tag{11}$$

여기서 μ는 $\boldsymbol{\Gamma_{VV}}$이 역행렬을 가질 수 있도록 하는 상수로, 본 논문에서는 10^{-4}으로 설정하였다. 이와 같은 빔형성기를 통하여 실제 환경에서 인접 배치된 스테레오 무지향성 마이크로폰을 활용하여 녹음 받은 음원의 강도차 극대화가 가능하다.

2. 강도차에 따른 음원 분리 기법

강도차에 따른 음원을 분리하기 위하여 먼저 아래의 수식처럼 azimuth-frequency plane을 생성한다[4].

$$\begin{aligned} AF_L(k,i) &= |\widehat{S_R}(k)| - g(i)|\widehat{S_L}(k)| \\ AF_R(k,i) &= |\widehat{S_L}(k)| - g(i)|\widehat{S_R}(k)| \end{aligned} \tag{12}$$

여기서 $\widehat{S_L}(k)$와 $\widehat{S_R}(k)$는 MVDR 빔형성기를 통하여 강도차가 극대화된 신호이며, $g(i)$는 gain scaling factor을 나타내며 아래의 수식과 같다.

$$g(i) = \frac{i}{\beta} \quad \text{for } 0 \le i \le \beta \tag{13}$$

여기서 β는 방위각에 대한 해상도를 나타내는 상수이며, β가 커질수록 방위각 해상도가 높아지지만, 그만큼 많은 연산을 요구하게 된다. 본 논문에서는 β를 90으로 설정하였다. <그림 2>는 방위각 -60도에서 재생한 백색 잡음을 5 cm 간격을 가진 무지향성 스테레오 마이크로폰으로 녹음받아 수식 (12)를 활용하여 생성된 azimuth-frequency plane의 실예를 나타낸다. <그림 2>에서 나타나듯이 azimuth가 20에서 40정도 범위에서 magnitude가 다른 범위보다 낮은 것을 확인 가능하다. 이는 백색 잡음이 azimuth가 20에서 40이 되는 범위 안에 존재한다는 것을 의미하며, 위의 과정을 통하여 생성된 azimuth-frequency plane를 다시 방위각에 따른 최소값과 최대값을 결정하고, 이에 따라 아래의 수식처럼 binary masking을 하게 된다.

$$\overline{AF_m}(k,i) = \begin{cases} AF_m^{\max}(k) - AF_m^{\min}(k) & if\ AF_m(k,i) = AF_m^{\min}(k) \\ 0, & otherwise \end{cases} \tag{14}$$

여기서 $AF_m^{\max}(k)$와 $AF_m^{\min}(k)$은 i에 따라 나타나는 값 중에서 최대값과 최소값이 가리키며, m은 L 혹은 R 채널을 의미한다. ADRess 음원 분리 기법은 원하는 방위각을 설정하여 그 방위각에 대한 소리만을 분리한다. 즉, 설정하는 방위각에 따라 수식 (13)에서 왼쪽 채널에 대한 binary masking이 진행될지, 혹은 오른쪽 채널에 대한 binary masking이 진행될지 결정하게 된다. 또한, 방위각과 거기에 따른 넓이 B를 설정하여 얼마만큼의 특정 방위각에 대한 정보를 가져올지 결정하게 된다. 즉, B가 높을수록 넓은 방위각에 대한 오디오 정보를 추출하게 되고, B가 작을수록 좁은 방위각에 대한 오디오 정보만을 분리하게 된다.

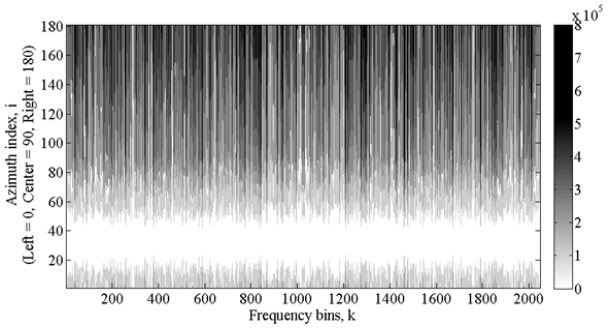


그림 2. 백색 잡음에 대한 azimuth-frequency plane
Fig. 2. Azimuth-frequency plane for the white noise.

$$|Y(k)| = \sum_{i=d_a-(B/2)}^{d_a+(B/2)} \overline{AF_m}(k,i) \tag{15}$$

여기서 d_a가 ADRess 음원 분리 기법에서 직접 설정해주어야 하는 방위각이며, 이 방위각 d_a와 B를 직접 조절하여 원하는 방향의 음원을 분리 가능하다. 마지막으로 수식 (15)로 획득한 magnitude 성분과 원음의 phase 성분을 가지고 최종적으로 음원 분리된 신호를 획득한다.

$$Y(k) = |Y(k)|\exp(j\widehat{S_m}(k)) \tag{16}$$

Ⅲ. 성능 평가

본 논문에서 제안한 음원 분리 기법의 성능을 검증하기 위하여 stereo audio source separation evaluation campaign (SASSEC)에서 제공하는 객관적 성능평가 툴을 활용하였다[8]. SASSEC에서는 분리된 음원을 총 네 가지 성분으로 분해하며, 수식은 아래와 같다.

$$y(n) = s(n) + e^{spat}(n) + e^{interf}(n) + e^{artif}(n) \tag{17}$$

여기서 $s(n)$은 음원 분리 기법을 통하여 얻고자 하는 이상적인 타겟 신호이며, $e^{spat}(n)$, $e^{interf}(n)$, $e^{artif}(n)$은 각각 spatial error, interference error, artifact error 성분을 뜻한다. 이 네가지 성분으로 분해하여 source-to-distortion ratio (SDR), source-to-interference ratio (SIR), and sources-to-artifacts ratio (SAR)이 결정 가능하다[8].

본 논문의 음원 분리 기법은 인접 배치된 스테레오

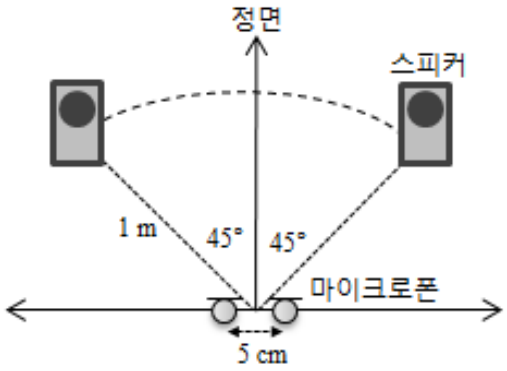


그림 3. 성능평가를 위한 마이크로폰과 스피커 배치도
Fig. 3. Configuration of microphones and loudspeakers for the performance evaluation.

표 1. 객관적 성능평가 SASSEC 결과
Table 1. The results of the SASSEC.

기법	SDR	SAR	SIR
Conventional	-2.76 dB	-8.02 dB	8.71 dB
Proposed	2.31 dB	5.64 dB	23.48 dB

무지향성 마이크로폰 환경을 고려한 음원 분리 기법이므로, 실제 무지향성 마이크로폰을 인접 배치하여 음원을 녹음받았다. <그림 3>은 본 논문에서 성능 평가를 위하여 마이크로폰과 스피커의 구성 환경을 나타낸다. 먼저, 스테레오 무지향성 마이크로폰의 간격은 5 cm로 배치하였고, 스피커와 마이크로폰의 간격이 1 m가 되도록 배치하였다. 스피커는 스테레오 마이크로폰의 중앙과 방위각이 -45도와 45도가 되도록 하였다. Sound quality assessment material (SQAM)에 포함되어 있는 여성 음성(track 49)과 남성 음성(track 50)을 각각 스피커로부터 재생하였다[9].

<표 1>은 객관적 성능평가 결과를 보여준다. 표에서 제안된(proposed) 기법은 본 논문에서 제안하는 음원 분리 기법을 나타내며, 기존(conventional) 기법은 MVDR 빔형성기를 사용하지 않고 강도차에 따른 음원 분리만을 적용했을 때의 방법을 나타낸다. 성능 평가결과, SIR이 8.71dB에서 23.48dB로 대폭 상승하였고, SDR과 SAR 측면에서도 기존 기법에 비해 우수하게 평가되었다.

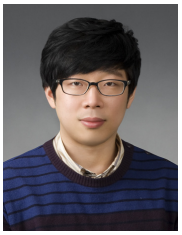
Ⅳ. 결 론

본 논문에서는 인접 배치된 스테레오 무지향성 마이크로폰 환경에서 강도차에 따라서 음원을 분리하는 기법을 제안하였다. 이를 위하여, 입력된 스테레오 오디오 신호의 강도차를 극대화하기 위하여 MVDR 빔형성기를 적용하였고, 이를 강도차에 따라서 음원 분리 기술을 적용하였다. 제안된 기법은 SASSEC에서 제공하는 객관적 성능평가 방법을 통하여 성능을 검증하였다. 검증한 결과, SDR, SAR, SIR 측면에서 모두 기존 기법에 비하여 우수한 성능을 보인 것을 확인하였다.

REFERENCES

[1] A. Hyvärinen, “Survey on independent component analysis,” *Neural Computing Surveys*, vol. 2, pp. 94-128, 1999.
[2] D. F. Rosenthal and H. G. Okuno, *Computational Auditory Scene Analysis*, LEA Publishers, Mahwah, NJ, 1998.
[3] P. Divenyi, *Speech Separation by Humans and Machines, Kluwer Academic Publishers*, Norwell, MA, 2005.
[4] D. Barry, B. Lawlor, and E. Coyle, “Sound source separation: azimuth discrimination and resyntesis,” in *Proc. of International. Conference on Digital Audio Effects (DAFX-04)*, pp. 1-5, Oct. 2004.
[5] 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.
[6] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, Springer, Berlin, Germany, 2001.
[7] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, Springer, Berlin, Germany, 2008.
[8] 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. of International Conference on Independent Component Analysis and Signal Separation*, pp. 552-559, Feb. 2007.
[9] EBU Technical Document 3253, *Sound Quality Assessment Material Recordings for Subjective Tests - Users’ Handbook for the EBU-SQAM Compact Disc*, 1988.

저 자 소 개



전 찬 준(학생회원)
2009년 한국기술대학교 전자공학과 학사 졸업.
2011년 광주과학기술원 정보통신공학부 석사 졸업.
2011년~현재 광주과학기술원 정보통신공학부 박사과정.
<주관심분야 : 오디오 신호처리, 3D 오디오>

정 석 희(학생회원)
2013년 충북대학교 정보통신공학과 학사 졸업.
2013년~현재 광주과학기술원 정보통신공학부 석사과정.
<주관심분야 : 오디오 신호처리, 3D 오디오>

김 홍 국(평생회원)-교신저자
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 정보통신공학부 교수.
<주관심분야 : 음성인식, 음성 및 오디오 신호처리, 3D 오디오>