시간-주파수 영역에서의 스테레오 사운드 분리기법

Stereo Sound Demixing Method in Time-Frequency Domain

이재은*, 김영문*, 임 찬*, 강현수**
중앙대학교 첨단영상대학원 영상공학과*, 충북대학교 전기전자컴퓨터공학부**

Jae-Eun Lee(jlee@wm.cau.ac.kr)*, Young-Moon Kim(kimoon@wm.cau.ac.kr)*,
Chan Lim(mj23cb@wm.cau.ac.kr)*, Hyun-Soo Kang(hskang@cbnu.ac.kr)**

요약

본 논문은 스테레오 사운드에서 합쳐지기 이전의 개별적인 사운드를 분리해내는 기법을 제안한다. 기존의 Degenerate Unmixing Estimation Technique (DUET) 알고리즘의 W-Disjoint Orthogonal 가정에 기반을 두고 있으며, Windowed-Fourier 변환을 사용하여 시간-주파수 영역에서 주요 프로세스를 수행한다. 제안된 방식은 패닝 인덱스의 거리차이에 따라 가중치를 준 마스크를 사용하는 기법과 양쪽 채널의 성분을 비교하여 바이너리 기반의 마스크를 사용하는 방식이다. 전자는 부드러운 분리 특성을 보여주며, 후자는 높은 분리 특성을 보여주었다. 마지막에 실험을 통해 기존의 방식과 제안된 방식을 비교함으로써, 제안된 방식이 기존 방식 보다 좋은 성능을 가지고 있음을 알아볼 것이다.

■ 중심어 : | 디믹싱 | 스테레오 사운드 | 블라인드 음원 분리 |

Abstract

This paper presents a new demixing method that separates each source from a stereo sound mixture. Under the W-Disjoint Orthogonal assumption in DUET(Degenerate Unmixing Estimation Technique) algorithm. The proposed method is mainly processed in time-frequency domain by using windowed-fourier transform. In this paper there are two main contributions: a weighted mask by panning index distances and a binary mask by comparing each channel value. The former has tender demixing characteristic, and the latter has stronger demixing characteristic. In experimental results, we will show that both masks produce more robust demixing than the existing demixing methods do.

■ keyword : | Demixing | Stereo Sound | Blind Source Separation |

I. 서 론

스테레오 방식은 2개의 채널을 사용하는 입체 음향 방식으로, 현재 가장 널리 사용되고 있는 믹싱(Mixing)/재생 방식이다. 스테레오 방식에서는 하나의 소스를 다른 크기로 양쪽 채널에 삽입하거나 한 쪽 채널에 지연을 두어 소스의 방향을 조절한다. 여러 개의 소스를 각각 다른 크기와 지연으로 믹싱해주면, 각각의 소스마다 다른 방향성을 줄 수 있다. 이는 스테레오 마이크를 사용한 라이브(Live) 녹음에서도 그대로 적용될 수 있

* 본 논문은 2006학년도 충북대학교 학술연구지원사업의 연구비지원에 의하여 연구되었습니다.

접수번호 : #0070418-002
접수일자 : 2007년 04월 18일
심사완료일 : 2007년 07월 18일
교신저자 : 강현수, e-mail : hskang@cbnu.ac.kr

는 모델로, 스테레오 마이크 자체가 채널 간 크기 차이와 지연 차이를 만들어내기 때문이다. 이와는 반대로 여러 개의 소스가 합쳐진 신호에서 합쳐지기 이전의 개별적인 소스들로 분리해내는 것을 디믹싱(Demixing)이라 한다. 본 논문은 여러 개의 소스가 스테레오로 믹싱되어있는 상태에서 합쳐지기 이전의 개별적인 소스들로 분리해내는 디믹싱 기법에 대해서 다루고 있다. 스테레오 음향 신호를 기반으로 믹싱 이전의 개별적인 소스들로 분리해내는 것은 다양한 활용성이 있다. 스테레오 신호 기반의 멀티채널 음향 시스템 구현[1], 음악 소스에서의 보컬 사운드 억제, 소스 개선, 개별적인 소스의 방향감 재조정[2], 가상 마이크로폰의 구현[3] 등 다양한 활용분야에 대한 연구도 있어오고 있다. 또한 다중 화자의 위치 확인, 개별적인 음향신호의 시각화 등에서도 사용이 가능할 것이다.

본 논문은 DUET 알고리즘에 기반하고 있으며, 음성과 음악 사운드 모두를 대상으로 하고 있다. 최종적으로 1)높은 분리도가 우선적인 기법과 2)부드러운 분리가 우선적인 기법 두 가지 접근 방향을 가지고 접근하고 있다.

II. DUET 알고리즘

본 장에서는 Blind Source Separation (BSS) 과 Degenerate Unmixing Estimation Techniques (DUET) 알고리즘에 대해서 알아본다.

2.1 Blind Source Separation 개요

여러 사운드가 합쳐진 사운드에서 개별적인 사운드를 분리하는 것은 칵테일 파티 효과를 해결하는 것과 유사하다 할 수 있다. 칵테일 파티 효과는 사람이 파티장과 같이 주위가 매우 시끄러운 환경에서도 들으려고 의식하는 소리를 선별적으로 들을 수 있지만, 이를 녹음하여 듣게 되면 듣지 못하는 현상을 말한다. 이는 인간의 청각이 가지고 있는 특수한 능력으로 기계적인 장비로 구현이 힘든 현상이다. 이러한 칵테일 파티 효과의 구현은 실제 인간의 지각 능력을 구현하는 것으로 신호처리, 음향뿐만이 아니라 인간 신경을 구현하고자 하는 신경 회로망(Neural Networks)과 같은 학문들과도 연관되어 다양한 연구가 이루어져오고 있다. 이와 같이 합쳐진 신호를 다시 나누는 것을 BSS, 즉 Blind Source Separation 또는 Blind Signal Separation[5]이라고 한다. 여기서 Blind는 원본 신호에 대한 정보가 없으며, 믹싱 된 신호에 대해서도 정보가 없다는 것을 뜻하는 것이다. 여기서 최종적으로 신호를 분리하는 과정을 디믹스(Demix) 또는 언믹스(Unmix)라는 용어를 사용하여 표현한다.

2.2 DUET(Degenerate Unmixing Estimation Technique)

Degenerate Unmixing Estimation Technique (DUET)는 시간-주파수 영역에서 이루어지는 대표적인 Blind Source Separation 기법이다[6]. 본 단락에서 DUET 알고리즘에 대해서 알아본다.

2.2.1 신호 모델

$s_0(t), s_1(t), s_2(t), \ldots, s_N(t)$와 같이 N개의 소스가 스테레오 방식으로 믹싱되어 있다면, 식(1)과 같은 신호 모델을 세울 수 있다.

$$x_1(t) = \sum_{i=0}^{N} s_i(t) + n_1(t)$$
$$x_2(t) = \sum_{i=0}^{N} \alpha_i s_i(t-\delta_i) + n_2(t) \quad (1)$$

여기서 $x_1(t)$는 왼쪽 채널의 신호, $x_2(t)$는 오른쪽 채널의 신호이며, $s_i(t)$는 i번째 소스, $n_1(t)$와 $n_2(t)$는 각각의 채널에 삽입된 잡음을 나타낸다. 여기서 α_i는 i번째 소스에 적용되어 양쪽 채널 간 크기 차이를 나타내는 매개 변수이며, δ_i는 i번째 소스에 적용되어 양쪽 채널 간의 지연을 나타내는 매개 변수이다.

이와 같은 신호모델은 스튜디오 믹싱뿐만 아니라 스테레오 마이크를 사용하여 소리를 입력받는데도 사용될 수 있다.

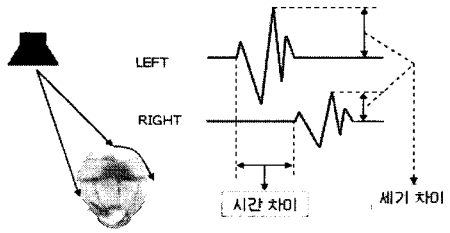


그림 1. 두 귀에 도달하는 소리의 세기차이와 시간차이

[그림 1]과 같이 특정 방향에서 소리가 전달되면 양쪽 귀에는 도달 거리에 따른 소리의 크기 차이와 지연이 발생 된다. 식(1)의 모델은 이 같은 상황을 그대로 반영하고 있다. DUET 알고리즘은 이와 같은 스테레오 신호 모델을 기반으로 잡음이 없다고 가정하여 모델을 단순화하여 사용하고 있다.

2.2.2 Windowed-Fourier Transform

DUET 알고리즘은 변환 $\mathcal{T}$로 Windowed-Fourier Transform을 사용한다. 이는 창 함수(Window Function)를 사용하여 푸리어 변환을 수행하는 것으로 Short-Time Fourier Transform이라고도 한다.

$$S_j^w(\tau,\omega) = F^w(s_j(t))(\omega,\tau) = \int_{-\infty}^{\infty} W(t-\tau)s_j(t)e^{-j\omega t}dt \quad (2)$$

Windowed-Fourier Transform은 시간 t에 대한 신호를 시간 프레임 τ와 주파수 ω에 대한 신호로 변환시키는 변환이다. 이는 시간 영역의 신호를 시간-주파수 영역으로 이동시키는 것으로, 변환의 결과로 각 시간 프레임 마다 주파수 성분을 얻을 수 있다. 이와 같이 시간-주파수 영역에서 프로세스를 수행하는 것은 음향 신호, 특히 음성신호의 경우 시간-주파수 영역에서 통계적인 독립성이 높기 때문이다[7].

2.2.3 W-Disjoint Orthogonal

DUET 알고리즘은 합쳐지기 이전의 신호들이 W-Disjoint Orthogonal(WDO)하다고 가정하고 접근하고 있다. 이는 DUET 알고리즘의 핵심으로 $i \neq j$라면 $s_i(t)$와 $s_j(t)$의 Windowed-Fourier Transform이 상호배반적(Disjoint)이라는 가정이다. 즉, 모든 시간과 모든 주파수에서 각각의 신호들이 서로 연관성이 없이 통계적으로 독립적이며, 따라서 합쳐진 신호의 모든 시간-주파수 성분은 하나의 신호와만 연관이 있다는 가정이다. 이를 수식으로 나타내면 식(3)과 같다.

$$S_i^w(\tau,\omega)\,S_j^w(\tau,\omega) = 0,\ \forall i \neq j,\ \forall \omega,\tau \quad (3)$$

이러한 가정은 실제 음향 신호에 완전하게 대응하지는 않지만, 연구를 통해 음성 신호의 경우에는 매우 적절하게 대응한다고 알려져 있다[7].

2.2.4 Amplitude-Delay Estimation[8]

2.2.3의 W-Disjoint Orthogonal 가정을 적용하면, 스테레오 신호 모델인 식(1)에서 α와 δ를 구할 수 있다. 식(1)의 Windowed-Fourier Transform은 식(4)로 표현될 수 있다.

$$\begin{bmatrix} X_1(\tau,\omega) \\ X_2(\tau,\omega) \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 1 \\ \alpha_1 e^{-j\omega\delta_1} & \cdots & \alpha_N e^{-j\omega\delta_N} \end{bmatrix} \begin{bmatrix} S_1(\tau,\omega) \\ \vdots \\ S_N(\tau,\omega) \end{bmatrix} \quad (4)$$

W-Disjoint Orthogonal 가정을 적용하면, 특정 시간 프레임 τ_n과 특정 주파수 ω_m에서의 식(4)는 식(5)로 정리될 수 있다. 가정에 의해 특정 i번째 소스를 제외한 모든 소스들이 제거되기 때문이다.

$$\begin{bmatrix} X_1(\tau,\omega) \\ X_2(\tau,\omega) \end{bmatrix} = \begin{bmatrix} 1 \\ \alpha_i e^{-j\omega\delta_i} \end{bmatrix} S_i(\tau,\omega) \quad (5)$$

식(5)로부터 연립방정식을 풀면 채널간의 상대적인 크기를 나타내는 매개 변수인 α_i와 지연을 나타내는 매개 변수인 δ_i를 식(11)과 같이 구할 수 있다. 여기서 $I(X)$는 복소수 X의 허수 성분을 나타낸다.

$$(\alpha_i, \delta_i) = \left(\left| \frac{X_2(\tau,\omega)}{X_1(\tau,\omega)} \right|, I\left(\log\left(\frac{X_1(\tau,\omega)}{X_2(\tau,\omega)} \right) \right) \Big/ \omega \right) \quad (6)$$

전체 시간 프레임과 주파수에 걸쳐 위의 과정을 거치면 모든 시간 프레임과 모든 주파수에 대해서 각각의 α 와 δ 를 구할 수 있다. 단, 지연 δ를 구하는데 있어서는 식(7)의 의미 있는 구간이 존재하는데, 이는 복소수가 가지고 있는 주기성에 기인한 것이다.

$$|\omega\delta_j| < \pi \tag{7}$$

2.2.5 W-Disjoint Orthogonal인 경우의 디믹싱 과정

식(6)를 통해 모든 시간-주파수의 (α,δ)를 구할 수 있으며, 이를 인덱스로 하여 동일한 (α,δ)를 가지는 시간-주파수 (τ,ω)의 집합 Λ 을 만들 수 있다. 신호들이 W-Disjoint Orthogonal에 완벽히 대응한다면 이와 같은 과정을 통해 얻은 집합 Λ 는 실제 신호의 개수인 N개 만큼 생성되며, 동일한 (α,δ)를 가지는 (τ,ω)는 하나의 신호를 구성하게 된다. 여기서 식(8)과 같은 마스크를 적용하면 특정 j번째 신호의 시간-주파수 성분을 분리해 낼 수 있다.

$$M_j(\tau,\omega)=\begin{cases}1 & (\tau,\omega)\in\Lambda_j\\ 0 & otherwise\end{cases} \tag{8}$$

$$S_j(\tau,\omega)=M_k(\tau,\omega,)X_1(\tau,\omega) \tag{9}$$

그리고 Inverse Fourier Transform하게 되면 $t=\tau$ 에서 원래 시간 영역 신호를 구할 수 있다.

$$s_j(t)=FT^{-1}(S_j(\tau,\omega)),t=\tau \tag{10}$$

종합적으로 DUET 알고리즘의 과정을 정리하면,

1) $x_1(t),x_2(t)$를 Windowed-Fourier Transform 하여 시간-주파수 영역으로 변환하고,
2) 모든 시간-주파수 (τ,ω)에 대한 (α,δ)를 구하고,
3) 동일한 (α,δ)를 가지는 (τ,ω)의 집합 Λ_i를 구성하고,
4) Λ_i에 대한 마스크를 구성하고,
5) 마스크를 통해 특정 i번째 신호의 시간-주파수 성분 $S_i(\alpha,\delta)$을 분리하고,
6) Inverse Fourier Transform하여 원래 시간 영역의 신호 $s_i(t)$를 얻는다.

위의 과정은 신호가 W-Disjoint Orthogonal 조건에 완벽히 만족할 경우에 사용할 수 있다. 하지만 실제의 음향 신호는 W-Disjoint Orthogonal 조건에 완벽히 만족하지 않기 때문에, 특정 시간-주파수 (τ,ω)에서 하나의 신호가 주요하여 다른 신호들이 미치는 영향이 적다는 가정을 하고 문제에 접근하여야 한다.

2.2.6 실제 음향 신호의 디믹싱 과정

실제 음향신호의 경우 위의 과정을 통해 α를 구하면 α값에 오차가 발생된다. 이는 실제 음향 신호가 W-Disjoint Orthogonal 가정을 완전히 만족하지 않기 때문으로, 이와 같은 이유로 2.2.5의 과정을 그대로 적용할 수는 없다. 따라서 전체 시간 프레임에 걸쳐 에너지 히스토그램을 구하고, 에너지가 높은 피크를 확인함으로써 실제로 사용된 α를 추측해내는 방법을 사용한다. 이 과정을 통해서 실제로 사용된 α,δ를 추측할 수는 있으나, 식(10)과 같은 형태의 마스크는 사용할 수 없으며, 따라서 다음과 같은 마스크를 사용한다. 여기서는 α를 중심으로 설명한다.

1) 특정 범위내의 α,δ를 가지는 성분들에 사각 윈도우(Rectangular Window) 적용[7]

이는 [그림 2]의 왼쪽과 같이 Δ라는 범위를 두고, 식(11)과 같이 분리하고자 하는 α_c를 중심으로 범위 Δ 안에 있는 α를 가지는 성분들에 바이너리 마스크를 적용하는 것이다.

$$M(\alpha,\Delta,\tau,\omega)=\begin{cases}1, & |\ln\alpha(\tau,\omega)-\ln\alpha_c|<\Delta/2\\ 0, & otherwise\end{cases} \tag{11}$$

이와 같이 가중치 없이 바이너리 마스크를 사용하는 방식은 수치적인 평가 척도를 기준으로 할 때는 좋은

분리 능력을 보여주지만, 음질 적인 측면에서의 왜곡이 심해진다. 또한, 이 방식은 상황에 따라 Δ를 정해주어야 하는 단점이 있다.

2) α,δ에 따라서 가우시안 윈도우(Gaussian Window) 적용[1]

이는 [그림 2]의 오른쪽 같이 분리하고자 하는 α_c를 중심으로 가우시안 윈도우를 적용하는 방식으로, 수식으로는 식(12)와 같다. 여기서 v는 왜곡을 줄이기 위한 기본 값(floor value)이며 ϵ는 윈도우의 두께를 결정하는 변수이다.

$$M(\alpha, \Delta, \tau, \omega) = v + (1-v)e^{-\frac{1}{2\epsilon^2}(|\alpha(\tau,\omega)-\alpha_c|)} \quad (12)$$

이와 같이 0과 1이 아닌 가중치가 적용된 마스크를 사용하는 것은 수치적인 평가 척도를 기준으로 할 때는 떨어지는 성능을 보여주지만, 음질적인 측면에서는 보다 부드러운 특성을 보여준다. 하지만 이 방식도 윈도우의 두께 ϵ를 상황에 따라 정해주어야 하는 단점이 있다.

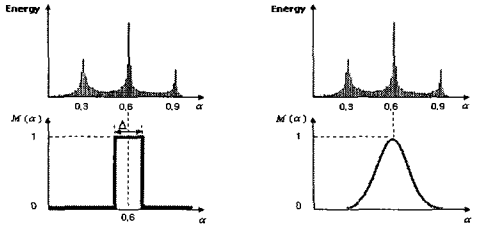


그림 2. 디믹싱 마스크 (왼쪽 위/오른쪽 위: α가 0.3, 0.6, 0.9로 믹싱 된 스테레오 소스의 에너지 히스토그램. 왼쪽 아래: α 0.6을 중심으로 적용된 Rectangular Mask, 오른쪽 아래: α 0.6을 중심으로 적용된 Gaussian Mask

3) Maximum Likelihood Estimation을 사용한 Binary Mask 적용[6]

분리하기 원하는 신호 이외의 신호를 가우시안 잡음(Gaussian Noise)형태라 가정하고, 각각의 신호들이 독립이라는 가정을 사용하면 식(13)과 같은 Likelihood Function을 얻을 수 있다. 여기서 식(14)와 같이 원하는 신호의 Likelihood 값이 다른 신호들에 비해 모두 큰 경우에는 1을 아닌 경우에는 0을 마스크로 사용하여 원하는 신호를 분리할 수 있다. 이와 같은 방식은 1)이나 2)와 같이 Δ, ϵ를 조절하지 않고 원하는 신호를 분리 할 수 있다는 장점이 있으며, 수치적인 평가 척도에서도 매우 좋은 성능 보여준다. 단, 몇 개의 신호가 합쳐져 있는지 정확하게 알고 있어야 좋은 결과를 기대할 수 있으며, 바이너리 마스크 사용에 따른 음질의 왜곡도 많이 발생하게 된다.

$$L_j(\tau,\omega) = \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}\|\alpha_j X_1(\tau,\omega) - X_2(\tau,\omega)\|^2/(1+\alpha_j^2)} \quad (13)$$

$$M(j,\tau,\omega) = \begin{cases} 1, & L_j(\tau,\omega) \geq L_i(\tau,\omega),\ \forall j \neq i \\ 0, & otherwise \end{cases} \quad (14)$$

2.2.7 Mask 평가 척도: WDO

해당 Mask의 성능을 평가하기 위해서 다음 두 가지 척도를 조합하여 사용한다.

1) Preserved Signal Ratio(PSR)

PSR은 해당 마스크가 추출하고자 하는 신호의 성분을 얼마나 유지 시키는가를 확인할 수 있는 척도로, j번째 신호에 대한 마스크 M의 PSR은 식(15)와 같다. 이는 해당 마스크가 j번째 신호의 에너지를 얼마나 유지시키지를 나타내는 것이다.

$$PSR_M = \frac{\|M(\tau,\omega)S_j(\tau,\omega)\|^2}{\|S_j(\tau,\omega)\|^2} \quad (15)$$

여기서 $\|f(x,y)\|^2 = \iint |f(x,y)|^2 dxdy$이며, 마스크의 범위가 $0 \leq M(\tau,\omega) \leq 1$이라면 PSR_M은 0부터 1사이의 값을 가지게 된다. PSR이 높을수록 신호가 많이 보존되었다고 할 수 있다.

2) Signal To Interference Ratio(SIR)

SIR은 해당 마스크가 간섭 신호의 성분을 얼마나 억제 시키는지를 나타내는 척도이다. 식(16)과 같이 $y_j(t)$가 j번째 신호를 제외한 신호의 합이라 하면, j번째 신호에 대한 마스크 M의 SIR은 식(17)과 같다.

$$y_j(t) = \sum_{\substack{k=1 \\ j \neq k}}^{N} s_k(t) \tag{16}$$

$$SIR_M = \frac{\|M(\tau,\omega)S_j(\tau,\omega)\|^2}{\|M(\tau,\omega)Y_j(\tau,\omega)\|^2} \tag{17}$$

여기서 $y_j(t)$는 간섭 신호의 합이 되며, SIR_M은 0부터 ∞사이의 값을 가지게 된다. SIR_M이 높을수록 간섭이 많이 제거되었다고 할 수 있다.

최종적으로 PSR과 SIR을 조합한 WDO라는 척도로 마스크의 성능을 평가한다. WDO는 식(18)과 같이 해당 마스크에 의해 유지된 신호의 에너지와 간섭 신호의 에너지 차이를 정규화(Normalizing)한 형태이다.

$$WDO_M = \frac{\|M(\tau,\omega)S_j(\tau,\omega)\|^2 - \|M(\tau,\omega)Y_j(\tau,\omega)\|^2}{\|S_j(\tau,\omega)\|^2} \tag{18}$$

$$= PSR_M - \frac{PSR_M}{SIR_M} \tag{19}$$

WDO가 높을수록 성능이 우수하다고 판단할 수 있다. WDO는 PSR이 1이고 SIR이 ∞일 때 1을 가지게 되며, 이 때가 가장 이상적인 경우이다.

III. 제안하는 디믹싱 기법

본 논문은 디믹싱 성능을 향상시키는 마스크를 제안하는 것에 중점을 두고 있으며, 이를 위해 식(20)과 같이 DUET 알고리즘의 신호 모델에서 시간적 지연을 제거시키고 모델을 단순화 하여 접근하였다. 여기서 양쪽 채널 간 크기차이를 발생시키는 매개 변수인 α를 패닝 인덱스(Panning Index)라 한다. 이와 같은 신호모델은 시간적 지연이 있는 실제 환경보다는 스튜디오 믹싱 환경에 더 적합한 모델이지만, 마스크의 성능을 평가하기에는 적절한 신호모델이 될 수 있다.

$$\begin{aligned} x_1(t) &= \sum_{i=0}^{N} \alpha_i s_i(t) \\ x_2(t) &= \sum_{i=0}^{N} (1-\alpha_i) s_i(t) \end{aligned} \tag{20}$$

DUET의 기본 가정인 W-Disjoint Orthogonal을 사용하고, $X(\tau,\omega)$가 $x(t)$의 Windowed-Fourier Transform이라 하면, 특정 시간-주파수 (τ,ω)에서의 패닝 인덱스 α는 식(21)에 의해서 구할 수 있다.

$$\alpha(\tau,\omega) = \frac{|X_1(\tau,\omega)|}{|X_1(\tau,\omega)| + |X_2(\tau,\omega)|} \tag{21}$$

본 논문에서는 위와 같이 DUET와 유사한 과정을 통해 패닝 인덱스를 구하였으며, 구해진 패닝 인덱스를 기반으로 신호를 분리할 수 있는 새로운 마스킹 기법을 제안하였다. 또한, 보다 유연성이 있는 평가 척도를 제시함으로써 보다 정교화 된 디믹싱 방향성을 제시하고자 한다.

3.1 평가 척도 - CR

앞서 제시된 평가 척도인 WDO는 마스크 자체의 평가에는 적합할 수 있지만, 실제로 분리된 신호와 원본 신호와의 차이를 확인하기에는 미흡한 점이 있다. 이는 마스크를 중심으로 접근하였기 때문이다. 따라서 본 논문에서는 최종 분리 신호의 성능 평가를 위해서 식(22)와 같은 평가 척도 Corrupt Rate(CR)를 도입하였다.

$$CR = \frac{1}{A}\sum_{\tau}\sum_{\omega}(\,|\,|S_j(\tau,\omega)| - |S_j'(\tau,\omega)|\,|) \tag{22}$$

여기서 $S_j(\tau,\omega)$는 j번째 원본 신호를, $S_j'(\tau,\omega)$은 j번째 분리된 신호를 뜻하며, A는 원본 신호의 전체 시간-주파수 성분의 크기 합, 즉 $\sum_{\tau}\sum_{\omega}|S(\tau,\omega)|$를 뜻한다. 이는 시간-주파수별로 분리된 신호와 원본 신호의

크기 차이를 원본 신호의 크기와 비교하고, 전체 크기 성분에서 해당 시간-주파수가 차지하는 비중을 가중치로 사용한 것이다. CR은 0부터 ∞까지의 값을 가지며, 0일 때 가장 작은 오차를 가져 가장 이상적인 성능을 가진다. 실제 주관적인 청취 테스트에서 CR이 낮을수록 더 높은 분리도와 부드러운 청감 특성을 보여주는 것을 확인하였으며, 테스트에 있어서 WDO와 함께 CR도 평가척도로 도입하였다.

3.2 Panning Index의 차이에 따른 선형 보간 마스크

제안하는 방식은 식(23)과 같은 형태의 마스크로 기존 2.2.6의 1)Rectangular Window를 사용하는 방식과 2)Gaussian Window를 사용하는 방식에 대응하는 마스크이다. 분리 능력보다는 분리된 사운드가 부드러운 청감특성을 가지도록 제안하였다. 분리를 원하는 신호의 패닝 인덱스와 간섭 신호의 패닝 인덱스, 그리고 해당 시간-주파수에서의 패닝 인덱스간의 차이를 보간하여 사용하는 형태로, 차이의 크기에 따라 패닝 인덱스 별로 가중치를 주는 것은 패닝 인덱스가 비슷할수록 간섭양이 크다는 사실에 기반하고 있다. 이와 같은 방식은 수치적인 평가 척도를 기준으로 할 때는 떨어지는 성능을 보여주나, 실제 듣기에는 훨씬 부드러운 특성을 보여준다. 단, 다른 신호의 간섭으로 인해 분리능력은 떨어지게 된다.

$$M_{PD}(\alpha_j,\tau,\omega)=\frac{\frac{1}{|\alpha_j-\alpha(\tau,\omega)|}}{\sum_{i=0}^{N}\frac{1}{|\alpha_i-\alpha(\tau,\omega)|}} \qquad (23)$$

만약 3개의 사운드가 패닝 인덱스 0.2, 0.4, 0.8로 믹싱되어있다면, [그림 3]과 같은 형태의 마스크가 구해진다. 기존의 Rectangular Window나 Gaussian Window보다 메인 로브가 줄고 사이드 로브가 넓어진 형태이다. 이는 기존 방식이 사이드 로브의 성분을 가져오지 않아 발생시킬 수 있는 왜곡을 줄여주는 효과를 기대할 수 있다. 여러 개의 신호가 합쳐짐에 따라 많이 왜곡된 패닝 인덱스가 구해질 가능성이 있기 때문이다. 단, 사이드 로브의 성분들로 인해 간섭 신호들이 섞이는 현상이 발생된다. 실제 청취 테스트에서도 다른 신호들이 섞이기는 하지만 기존의 방식보다 매우 부드러운 청감 특성을 보여주었다.

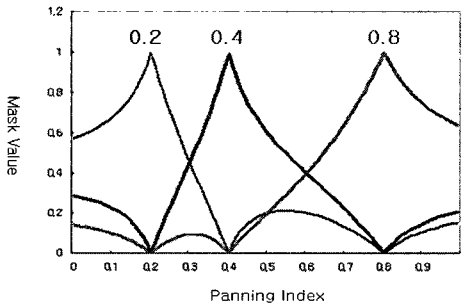


그림 3. 패닝 인덱스별 마스크 값(패닝 인덱스가 0.2, 0.4, 0.8 일 때)

제안된 마스크는 실제로 합쳐진 신호의 개수를 모르는 상태에서도 활용이 가능하며, 이를 업믹스(Upmix)에 쉽게 응용할 수 있다. 실제 신호의 개수와 패닝 인덱스를 모르는 상태에서 가상으로 0.4, 0.5, 0.6의 패닝 인덱스를 가진다고 가정하고, 각각의 패닝 인덱스별로 제안된 마스크를 적용하여 신호를 추출하면, 2채널의 신호에서 3채널의 신호를 얻을 수 있다. 여기서 가상의 패닝 인덱스의 값을 조절하여 음장의 범위와 방향감 등을 제어할 수 있다. 또한, 제안된 마스크는 그 합이 1이라는 점에서 전체 신호의 크기를 왜곡하지 않고 업믹스를 수행 할 수 있다는 장점도 가지고 있다.

3.3 Least Difference 마스크

여기서 제안하는 방식은 식(25)와 같은 형태로 기존 2.2.6의 3) Maximum Likelihood를 사용하는 방식에 대응하는 방식이다. 제안 하는 방식은 임의의 수 A가 $X_1=\alpha\times A$, $X_2=(1-\alpha)\times A$로 믹싱 되어 있을 때 $\frac{(1-\alpha)}{\alpha}\times X_1-X_2=0$이라는 기본적인 수학에 바탕을 두고 있다. 식(24)와 같이 $LD(\tau,\omega)$를 구하고, 가장 작은 $LD(\tau,\omega)$를 가질 때 해당 신호일 가능성이 가장 크다고 가정하여 $LD(\tau,\omega)$가 가장 작을 경우는 1을 마스크

값으로 설정하고, 그렇지 않을 경우는 작은 값을 설정하여 왜곡을 줄이려고 시도하였다.

$$LD_j(\tau,\omega)=\frac{(1-\alpha_j)}{\alpha_j}\left(\left|X_1(\tau,\omega)\right|-\left|X_2(\tau,\omega)\right|\right) \tag{24}$$

$$M(j,\tau,\omega)= \begin{cases} 1, LD_j(\tau,\omega)\le LD_i(\tau,\omega),\ \forall j\neq i \\ a\times EnergyRate_j(\tau)\times M_{PD}(\tau,\omega), o.w. \end{cases} \tag{25}$$

여기서 a는 가중치를 조절하기 위한 값이며, $EnergyRate_j(\tau)$는 해당 시간 프레임에서 j번째 신호가 차지하는 비중이고, $M_{PD}(\tau,\omega)$는 앞의 3.2에서 제안된 마스크이다. 가능성이 가장 크지 않은 경우에도 작은 값을 곱해주는 이유는 실제로 신호가 작은 값일지라도 모든 주파수에 걸쳐 성분을 가지고 있기 때문에 이를 보완하여 왜곡을 줄이기 위해서이다. 이를 위해 해당 시간 프레임에서 신호가 차지하는 비중과 패닝 인덱스에 따라 보간 한 마스크를 곱하여 사용하였다. $EnergyRate_j(\tau)$를 사용하는 이유는 해당 프레임에서의 에너지 비중이 해당 프레임에서의 마스크 값에 영향을 줄 가능성이 크기 때문이며, 패닝 인덱스에 따라 보간 한 값을 사용하는 것은 패닝 인덱스에 따른 간섭양을 반영하기 위해서이다. $EnergyRate_j(\tau)$는 추출한 모든 패닝 인덱스에 걸쳐서 얇은 두께의 가우시안 윈도우를 사용하여 추출한 에너지를 사용하여 구할 수 있다. [그림 4]는 4개의 음성 신호가 믹싱 되어있을 때 실제의 에너지 분포와 식(27)을 통해서 추출한 에너지 분포를 비교하여 보여주고 있다. 두 차이가 적은 것으로 보아 에너지 분포를 추출하여 활용할 수 있음을 확인할 수 있다.

$$Energy_i(\tau)=\left(\left|X_1(\tau,\omega)+X_2(\tau,\omega)\right|^2\right)\times e^{-\frac{1}{2\epsilon^2}(|\alpha(\tau,\omega)-\alpha_i|)} \tag{26}$$

$$EnergyRate_j(\tau)=\frac{Energy_j(\tau)}{\sum_{i=0}^{N}Energy_i(\tau)} \tag{27}$$

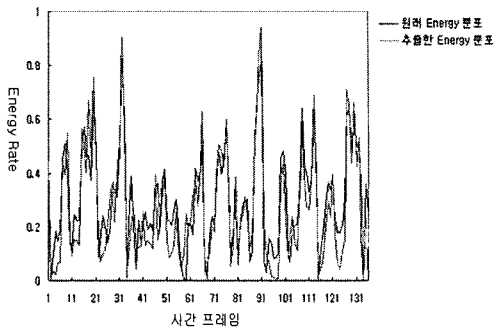


그림 4. 에너지 분포(4개의 음성 신호가 각각 패닝 인덱스 0.37, 0.43, 0.55, 0.64로 믹싱 된 스테레오 신호에서 패닝 인덱스 0.55인 신호가 전체 에너지 중 차지하는 비중을 식(27)을 사용하여 추출하고 실제 비중과 비교

IV. 실험 결과 및 분석

4.1 실험 환경

본 논문에서는 음성과 음악 소스 모두에 대해 테스트를 실시하였다. 음성 신호는 TIMIT 데이터베이스에서 무작위로 선별된 신호를 사용하였으며, 샘플링레이트는 16000Hz, 양자화비트는 16Bit인 파일을 사용하였다. 음악 신호는 모노로 되어있는 악기 및 보컬 신호를 사용하였으며, 샘플링레이트는 44100Hz, 양자화비트는 16Bit인 파일을 사용하였다. Windowed-Fourier Transform을 위하여 1024 샘플을 하나의 구간으로 취하여 1024 Point FFT를 수행하였으며, 윈도우는 Hamming 윈도우를 사용하고 50%의 오버랩 구간을 두었다.

4.2 Least Difference 방식의 디믹싱 결과

다수의 음성신호를 믹싱하여, 기존의 Maximum Likelihood를 사용한 방식과 제안 한 Least Difference 방식으로 Demixing 하여 결과를 비교하였다. [표 1]의 결과를 보면 제안된 방식이 CR 측면에서 더 우수한 성능을 보여주며, WDO 측면에서도 더 우수한 성능을 보여준다는 것을 확인할 수 있다.

표 1. Maximum Likelihood 방식과 Least Difference 방식 비교: 음성 신호. ML: Maximum Likelihood 방식, LD: Least Difference 방식

Mixing Source	Panning Index	CR		WDO	
		ML	LD	ML	LD
3개	0.50	0.354	0.330	0.851	0.853
	0.45	0.466	0.437	0.868	0.868
	0.60	0.330	0.324	0.922	0.924
4개	0.50	0.561	0.538	0.766	0.767
	0.45	0.532	0.495	0.809	0.827
	0.40	0.584	0.556	0.746	0.777
	0.58	0.481	0.487	0.788	0.777
5개	0.50	1.080	1.052	-0.480	-0.484
	0.45	0.667	0.635	0.577	0.587
	0.65	0.418	0.434	0.849	0.831
	0.40	0.711	0.687	0.609	0.622
	0.56	0.754	0.742	0.469	0.429

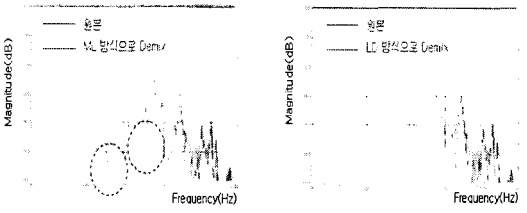


그림 5. 원본 신호와 분리한 신호의 주파수 스펙트럼 비교. 왼쪽: Maximum Likelihood(ML) 방식 사용, 오른쪽: 제안된 Least Difference(LD) 방식 사용.

[그림 5]는 여러 개의 음성이 믹싱 된 스테레오 신호에서 기존 Maximum Likelihood(ML) 방식과 제안된 Least Difference(LD) 방식으로 디믹싱 한 신호의 특정 시간 프레임에서의 주파수 스펙트럼을 원본과 비교한 것이다. 그림과 같이 ML 방식에서 차이가 나는 부분들이 LD 방식에서 많이 완화되었음을 확인 할 수 있다. [그림 6]은 시간 파형의 크기를 원본과 비교한 것으로, 마찬가지로 제안된 LD 방식이 왜곡을 줄여주는 것을 확인할 수 있다.

[표 2]는 음악신호를 분리한 것으로, 마찬가지로 LD 방식이 더 우수한 성능을 보여주고 있음을 확인 할 수 있다. 단, 맨 아래 항목은 동일한 음을 가지는 음악 소스들을 믹싱하여 분리한 것으로, 수치상으로 분리도가 매우 떨어진다는 것을 확인할 수 있다. 이는 기본 주파수가 동일하여 가정인 W-Disjoint Orthogonal에 많이 위배되기 때문으로 판단된다.

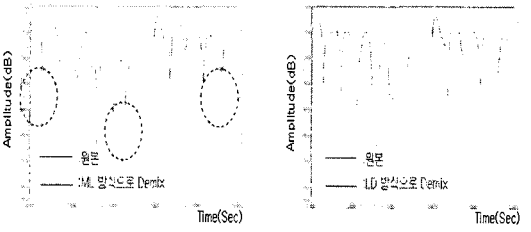


그림 6. 원본 신호와 분리한 신호의 시간 파형비교. 왼쪽: Maximum Likelihood(ML) 방식 사용, 오른쪽: 제안된 Least Difference(LD) 방식 사용

표 2. Maximum Likelihood 방식과 Least Difference 방식 비교: 음악 신호,ML: Maximum Likelihood 방식, LD: Least Difference 방식

Mixing Source	Panning Index	CR		WDO	
		ML	LD	ML	LD
3개	0.50	0.200	0.185	0.983	0.983
	0.45	0.250	0.238	0.945	0.948
	0.57	0.209	0.201	0.987	0.986
4개	0.50	0.371	0.346	0.929	0.931
	0.44	0.361	0.341	0.850	0.861
	0.54	0.504	0.484	0.853	0.859
	0.59	0.258	0.248	0.903	0.904
3개	0.50	0.694	0.631	0.351	0.299
	0.30	0.745	0.714	0.012	-0.037
	0.70	0.562	0.629	0.378	0.286

4.3 패닝 인덱스의 거리에 따른 선형 보간 마스크

다양한 소스를 믹싱하여 3.2에서 제안된 패닝 인덱스의 차이에 따른 선형 보간 마스크 방식을 사용하여 디믹싱을 수행한 후 기존의 방식과 비교하며 주관적인 청취 테스트를 실시하였다. 제안된 방식은 넓은 영역의 신호를 가져오기 때문에 다른 방식들에 비해서 부드러운 청감 특성을 보여주었다. 하지만 WDO나 CR과 같은 평가 척도에 있어서는 앞에서 살펴본 Maximum Likelihood나 Least Difference보다 떨어지는 성능을 보여주었으며, 실제로 다른 간섭 신호들이 섞이는 특성을 보여주었다. 단, 비슷한 접근이라 할 수 있는 가우시안 마스크를 사용할 때 보다는 좋은 성능을 보여주었다. 가우시안 마스크의 두께를 조절하는 파라미터인 ϵ를 조절하며 비교 평가를 해보았는데, 제안된 마스크가 비슷한 수준의 분리능력에서 더 부드럽게 들리는 특성을 보여주었다. 이는 앞에서 언급한 것과 같이 대부분의

신호가 모든 주파수 성분을 가지고 있으며, 제안된 방식은 사이드 로브 쪽의 성분들도 가져오기 때문으로 판단된다.

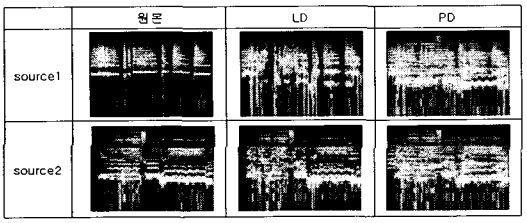


그림 7. Maximum Likelihood 방식과 패닝 인덱스를 사용하는 방식 비교: 음악 신호 (ML: Maximum Likelihood 방식, PD: 패닝 인덱스의 차이에 따른 방식, x축: 시간(0-3 sec), y축: 주파수 (20Hz~20kHz))

[그림 7]은 4개의 소스가 스테레오로 합쳐진 소스에서 두 개의의 소스를 3.3의 Least Difference 방식과 3.2의 패닝 인덱스에 따른 방식으로 분리한 결과와 원본의 스펙트로그램을 비교한 것이다. 3.3의 Least Difference 방식에 비해 3.2의 패닝 인덱스의 차에 따른 선형 보간 마스크 방식이 대부분의 주파수에서 더 많은 에너지를 가지고 있으며, Least Difference 방식이 원본에 더 가까운 형태를 보여주는 것을 확인 할 수 있다. 이와 같은 이유로 수치적인 평가척도에서는 Least Difference가 더 나은 결과를 보여주게 되지만, 음질적인 왜곡 측면에서는 Panning Index의 거리에 따른 선형 보간 마스크를 사용하는 것이 더 좋은 결과가 나오게 된다.

제안된 방식의 부드러운 분리 능력은 자연스럽게 들려야 하는 음악 신호의 디믹싱에 적절할 것으로 판단된다. 이러한 특징을 고려하여 3.2에서 언급된 업믹스 방식을 사용하여, 스테레오 신호에 가상의 패닝 인덱스를 설정하고 3채널 및 4채널로 업믹스 하는 테스트도 실시하여 좋은 결과를 얻을 수 있었다.

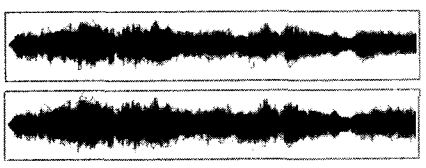


그림 8. 믹싱 된 스테레오 신호

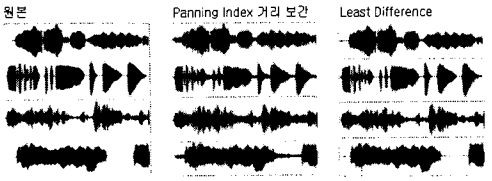


그림 9. 원본과 디믹싱 된 신호의 시간파형

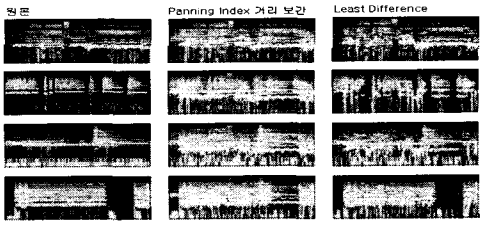


그림 10. 원본과 디믹싱 된 신호의 스펙트로그램

4.4 최종 결과

앞에서 살펴본 것과 같이 제안된 방식들은 기존 방식보다 좋은 성능을 보여주었다. 높은 분리도가 요구 된다면 Least Difference 방식을, 부드러운 청감 특성이 필요하다면 패닝 인덱스의 차이에 따른 보간 방식을 사용하는 것이 적절할 것이다.

[그림 8]은 4개의 음악 신호가 스테레오로 믹싱 되어 있는 신호이다. [그림 9]와 [그림 10]은 [그림 8]의 스테레오 신호를 제안된 방식으로 디믹싱 한 결과의 시간파형과 스펙트로그램이다. 패닝 인덱스의 거리를 사용하는 방식이 Least Difference 방식보다 간섭 성분이 덜 제거되었다는 것을 확인 할 수 있다. 하지만, 앞에서 살펴본 것과 같이 실제 청취 테스트에서는 패닝 인덱스의 차이를 사용하는 방식이 더 부드러운 청감 특성을 보여주었다.

[그림 11]과 [그림 12]는 4개의 음악 신호가 스테레오로 믹싱 되어있는 신호를 다양한 방식으로 디믹싱하고, 시간 프레임별 WDO와 CR을 나타낸 것이다. 여기서 0dB는 식(28)과 같이 분리하고자 하는 신호의 성분이 전체 성분의 절반 이상이 되면 마스크 값으로 1을 설정하고, 아닌 경우에는 0을 설정하는 것으로 DUET 알고리즘에서 레퍼런스로 사용된 것이다. 여기서 $Y_j(\tau,\omega)$는 식(16)과 같이 간섭 신호의 시간-주파수

성분을 뜻한다. Weight는 식(29)와 같이 전체 성분에서 분리하고자 하는 신호의 성분이 차지하는 비중을 마스크로 사용하는 것으로 본 논문에서 레퍼런스로 사용한 것이다. 이 두 레퍼런스 마스크는 믹싱 이전의 신호를 알고 있는 상태에서만 마스크를 구할 수 있으므로, Blind 방식이 아니며 레퍼런스로만 사용할 수 있다. LD는 Least Difference 방식을 뜻하며, PD는 패닝 인덱스의 차이에 따른 마스크를 뜻한다.

$$M(\alpha,\Delta,\tau,\omega)=\begin{cases}1, & 20log\left(\dfrac{|S_j(\tau,\omega)|}{|Y_j(\tau,\omega)|}\right)\geq 0\\ 0, & otherwise\end{cases} \quad (28)$$

$$M(\alpha,\Delta,\tau,\omega)=\frac{|S_j(\tau,\omega)|}{|X_1(\tau,\omega)+X_2(\tau,\omega)|} \quad (29)$$

여기서 주목해야 할 것은 크게 두 가지가 있다. 첫째, Weight 마스크가 WDO의 측면에서는 0 dB 마스크보다 낮은 성능을 보여주지만, CR의 측면에서는 매우 안정적인 형태를 보여준다는 것이다. 그리고 실제 청취 테스트에서는 Weight 마스크가 0 dB 마스크보다 더 좋은 특성을 보여 주었다. 이는 본 논문에서 사용한 척도인 CR이 적절한 척도임을 입증하는 것이며, 바이너리 마스크보다는 마스크에 가중치를 주는 것이 더 좋은 접근방식이라는 것을 나타낸다.

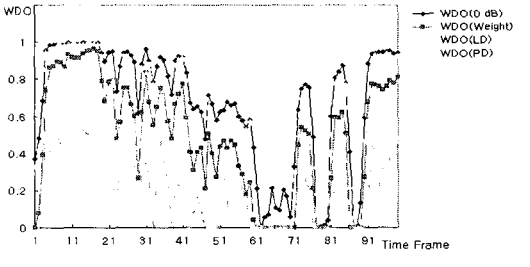


그림 11. 시간 프레임별 WDO 비교

둘째, 레퍼런스로 사용된 0 dB나 Weight 마스크에 비해 제안된 방식의 성능이 떨어지는 것을 확인할 수 있다. 이는 추가적인 연구를 통해서 성능 개선의 여지가 있음을 보여주는 것이다.

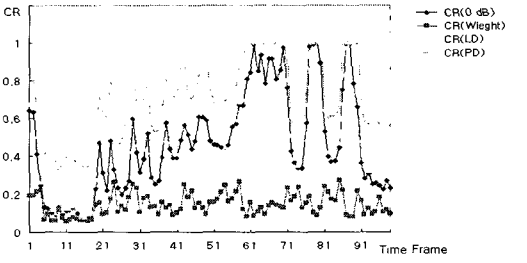


그림 12. 시간 프레임별 CR 비교

V. 결 론

본 논문에서는 스테레오 음향 신호에서 각각의 소스를 분리해내는 기법에 대해서 연구하였다. 패닝 인덱스의 차이에 따라 보간 한 값을 마스크로 사용하여 보다 부드러운 분리가 가능한 기법을 제안하였으며, Least Difference 방식을 제안하여 보다 분리도가 높은 디믹싱 기법을 제안하였다. 그리고 이를 다양한 테스트를 통해서 확인하였다. 하지만 이상적인 레퍼런스 마스크와 제안된 방식간의 성능 차이가 있었음을 확인하였기에, 좀 더 개선된 마스킹 기법을 찾아내는 것이 추후 과제가 될 것이다. 앞선 실험 결과와 같이 적절한 가중치가 적용된 마스크를 적용함으로써 성능향상을 기대할 수 있다. 그리고 제안 된 방식을 활용할 수 있는 분야에 대해서도 더 많은 고찰이 필요하다.

참 고 문 헌

[1] C. Avendano and J. M. Jot, "Frequency-Domain Techniques for Stereo to Multichannel Upmix," Proceedings of AES 22nd International Conference on Virtual Synthetic and Entertainment Audio, pp.121-130, 2002.

[2] C. Avendano, "Frequency-Domain Source Identification and Manipulation in Stereo Mixes for Enhancement, Suppression and Re-Panning Applications," Proceedings of IEEE Workshop on Application of Signal Processing to Audio and

Acoustics, pp.55-58, 2003.

[3] A. Radke and S. Richard, "Audio Interpolation for Virtual Audio Synthesis," Proceedings of AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio, pp.51-57, 2002.

[4] A. Hyarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, Wiley-Interscience, 2001.

[5] J. F. Cardoso, "Blind Signal Separation: Statistical Principles," Proceedings of the IEEE, Vol.86, No.10, pp.2009-2025, 1998.

[6] 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, 2004.

[7] S. Rickard and O. Yilmaz, "On The Approximate W-Disjoint Orthogonality Of Speech," Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vol.3, pp.3049-3052, 2002.

[8] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind Separation of Disjoint Orthogonal Signals: Demixing N Sources from 2 Mixtures," Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP), Vol.5, pp.2985-2988, 2000.

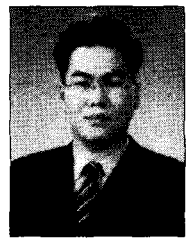
저 자 소 개

이 재 은(Jae-Eun Lee) 정회원

- 2001년 2월 : 중앙대학교 기계공학부 졸업(공학사)
- 2001년 ~ 2003년 : LG CNS 연구원
- 2003년 3월 ~ 2005년 2월 : 중앙대학교 첨단영상대학원 영상공학과 졸업(공학석사)

<관심분야> : 사운드, 영상처리

김 영 문(Young-Moon Kim) 정회원



- 2003년 2월 : 중앙대학교 컴퓨터공학과 졸업(공학사)
- 2003년 3월 ~ 2005년 2월 : 중앙대학교 첨단영상대학원 영상공학과 졸업(공학석사)

<관심분야> : 영상처리, 영상통신

임 찬(Chan Lim) 정회원

- 2000년 2월 : 아주대학교 전자공학과 졸업(공학사)
- 2000년 ~ 2002년 : LG전자 연구원
- 2003년 9월 ~ 2005년 8월 : 중앙대학교 첨단영상대학원 영상공학과 졸업(공학석사)

<관심분야> : 영상처리, 영상통신

강 현 수(Hyun-Soo Kang) 종신회원

- 1999년 2월 : KAIST 전기및전자공학과 졸업(공학박사)
- 1999년 ~ 2001년 : 현대전자 과장
- 2001년 ~ 2002년 : 한국전자통신연구원 선임연구원
- 2002년 ~ 2004년 : 중앙대학교 첨단영상대학원 영상공학과 조교수
- 2005년 3월 ~ 현재 : 충북대학교 전기전자컴퓨터공학부 부교수

<관심분야> : 영상처리, 영상부호화, 컨텐츠보호기술, 사운드 등