[5A-73]

스테레오 음향 신호의 새로운 Unmixing 기법

이재은, 강현수
중앙대학교 첨단영상대학원 영상공학과
jlee@wm.cau.ac.kr, hskang@cau.ac.kr

New Unmixing Method of Stereo Mixed Sound Signal

JaeEun Lee, Hyun-Soo Kang
Graduate School of AIM, Chung-Ang University

Abstract

In this paper, we propose an effective unmixing method to separate each source from a stereo mixed signal. The proposed method uses Windowed-Fourier Transform with the assumption of W-disjoint orthogonality introduced in the degenerate unmixing estimation technique (DUET) algorithm. We simplify the DUET algorithm by removing the delay factor between channels, and adopt other factors like phase difference for unmixing sources. As a result, the proposed method yields more improved unmixing ability, which will be verified by experimental results. Conclusively, the proposed method is more suitable for an amplitude-paned stereo mixture.

I. Introduction

스테레오 방식은 2 개의 채널을 사용하는 입체 음향 방식으로, 현재 가장 널리 사용되고 있는 믹싱/재생 방식이다. 스테레오 믹싱 방식에서는 하나의 소스를 다른 크기로 양쪽 채널에 삽입함으로써 재생시 방향감을 구현 할 수 있다. 이와 같은 방식을 Amplitude-Pan 이라 한다.

스테레오 신호에서 믹싱 되기 이전의 개별적인 음향 소스들을 추출하는 방식에 관한 다양한 연구들이 있어 왔다[1,4]. 또한, 이를 활용하여 스테레오 신호를 멀티 채널 환경에서 효과적으로 재생하는 방식[2], 가상의 마이크로폰을 구현하는 방식[3]등 구체적인 활용방안에 대한 연구도 활발하게 이루어져오고 있다. 위에 언급된 연구들은 DUET(Degenerate Unmixing Estimation Technique) 방식에서 제안된 개념인 음향 신호가 W-Disjoint Orthogonal 하다는 가정에 기반하고 있으며, 주요 프로세싱은 시간-주파수 영역에서 이루어진다. 이 가정은 믹싱 되기 이전의 각각의 소스들이 모든 시간-주파수 영역에서 겹쳐지지 않는 다는 가정이다[1]. 이러한 가정을 사용하여 음향 신호를 Windowed-Fourier Transform 을 사용해 시간-주파수 영역으로 변환시켜 신호를 추출하고, 다시 Inverse-Fourier Transform 하여 시간 영역으로 변환시킴으로써 믹싱되기 이전의 개별적인 소스들을 얻을 수 있다.

그렇다면, 과연 가장 기본적인 가정인 W-Disjoint Orthogonal 이라는 개념이 음향신호에 적적할게 적용될 수 있을까? 여러 연구결과에 따르면 W-Disjoint Orthogonal 이라는 가정이 제한된 숫자의 음성소스의 경우에는 비교적 잘 대응될 수 있으나[3,5], 일반적인 음악소스에 적용하기에는 문제점을 가지고 있다고 알려져 있다[6]. 이 논문은 기존의 연구들과 같이 W-Disjoint Orthogonal 이라는 가정에 기반을 두고 있으나, 양쪽 채널간의 위상차이를 고려하여 사용된 가정이 가지고 있는 오류를 완화시키는 과정과 추출시 사용되는 Weight factor 를 제안하여 보다 효과적인 Unmixing 기법을 구현하였다.

II. Signal Model and Fundamental Framework

2.1 Signal Model

이 논문에서 대상으로하는 신호는 Amplitude-Panning 방식으로 스튜디오 믹싱된 음향 신호에 한정하였으며, 다음과 같이 신호 모델을 세우고 단순화 시켰다.

N 개의 소스가 스테레오로 믹스되었다고 하면 다음과 같이 신호 모델을 세울 수 있다.

$$x_1(t) = \sum_{j=1}^{N} \alpha_j s_j(t) + n_1(t)$$
$$x_2(t) = \sum_{j=1}^{N} (1-\alpha_j) s_j(t-\delta_j) + n_2(t) \quad (1)$$

여기서 $s_j(t)$ 는 각각의 원래 신호들, $x_1(t)$ 는 믹싱된 왼쪽 채널의 신호, $x_2(t)$ 는 믹싱된 오른쪽 채널의 신호, α_j 는 얼마나 패닝이 되었는지를 나타내는 Panning-Coefficient, δ_j 는 왼쪽 채널에 비해서 오른쪽 채널이 얼마나 지연(Delay) 되었는지를 나타내는 Delay-Coefficient, 그리고 $n_1(t)$ 과 $n_2(t)$ 는 각각의 채널에 삽입된 노이즈이다.

식(1)의 모델은 양 채널간 지연(Delay)를 고려한 모