논문 2010-47TC-8-11

스테레오 음악 신호에서의 보컬 음원 분리를 위한 통합 알고리즘

(A Unified Method for Vocal Source Separation From Stereophonic Music Signals)

김 민 제*, 장 인 선*, 강 경 옥*

(Minje Kim, Inseon Jang, and Kyeongok Kang)

요 약

본 논문에서는 스테레오 형식의 음악 신호에서 가창 신호와 같은 음원을 분리하기 위한 통합 알고리즘을 제시한다. 스테레오 형식의 음악 신호에서 특정한 악기 음원을 분리하는 문제는, 획득한 음악 신호가 다양한 악기들이 동시에 연주되는 혼합 신호라는 점을 고려하고, 각각의 악기를 음원이라고 가정할 때, 획득한 혼합 신호의 개수가 음원의 개수보다 적은 비결정(underdetermined) 환경에서의 음원 분리 문제가 된다. 비결정 환경에서는 신호가 혼합되는 공간에 대한 가정을 기반하는 전통적 음원 분리 방식을 적용하기 힘들며, 목표 음원의 특정한 특성을 활용하여 추출하게 된다. 본 논문에서 제안하는 통합 알고리즘은 이종의 특성을 활용하는 음악 음원 분리 알고리즘들을 유기적으로 통합하는 구조이며, 구체적으로는 가창 신호와 같은 특정한 음원 추출을 위해 주로 사용되어 왔던 스테레오 채널 정보를 활용하는 방식과, 모노 혼합 신호에서 두드러지는 음원의 음정을 이용하여 음원을 추출하는 두 가지 방식을 통합하는 것을 목표로 한다. 본 논문에서 제안하는 구조는 각각의 음악 음원 분리 알고리즘이 가지고 있는 고유의 약점을 해소함으로써, 목표 음원의 복원 신호가 통합 과정에 의해 향상될 수 있다는 강점이 있으며, 그것을 실제 상업 음악 콘텐츠를 대상으로 한 실험을 통해 검증한다.

Abstract

A unified method for separating musical sources, singing voice for example, from stereophonic mixtures is provided. We usually have two observed signals in stereophonic music contents, where more than two instruments are played together. If we regard each instrument as source, this problem becomes an underdetermined source separation problem and cannot be solved by conventional methods, which infers the spatial environment of the downmixing process happens. Instead, source-specific information has been exploited to recover a particular instrumental source. This paper provides a unifying structure consists of heterogenious ad-hoc separate algorithms, which are designed for separating vocal sources using stereophonic channel information and dominant pitch information of the sources, respectively. Experiments on real world music contents show that the proposed unification can neutralize the drawbacks of the two ad-hoc separation algorithms and finally enhance the separation results.

Keywords : 음원분리, 음악음원분리, 센터채널추출, 비음성 행렬 인수분해, 객체기반 오디오 서비스

I. 서 론

음악 음원 분리는 여러 악기로부터 연주된 신호가 혼합되어 있는 상황에서 특정 악기의 연주만을 분리해내는 것을 목표로 하는 연구 분야이다. 기본적으로 가장

* 정회원, 한국전자통신연구원
(Electronics and Telecommunications Research Institute)

※ 본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2010년도 문화콘텐츠산업기술지원사업의 연구결과로 수행되었음.

접수일자: 2010년7월6일, 수정완료일: 2010년8월6일

유망한 응용 분야로는 원본 음악에서 가창 음원(vocal source)만을 분리해서 제거한 다음 남은 신호를 고품질 노래방 반주 신호로 활용하는 것이다. 이는 기존의 MIDI(Musical Instrument Digital Interface)와 가상 악기를 활용한 노래방 반주에 비해 보다 원곡에 가까운 반주를 사용자에게 제공할 수 있다는 장점이 있다. 또한, 최근 각광받고 있는 IM AF(Interactive Music Application Format)와 같은 관련 표준 기술[1~2]과 Music 2.0을 비롯한 객체기반 오디오 서비스의 일종인 멀티트랙 음악 서비스에서는 악기별 트랙을 사용자에게 제공함으로써 사용자가 전체적인 음량 뿐 아니라 악기별 음량을 제어할 수 있도록 해준다. 이러한 서비스를 위해서는 역시 악기별로 분리되어 있는 음원 신호가 필요하며, 이 때 가창 음원을 포함해 보다 다양한 음원의 필요성이 함께 대두된다.

또한 5.1 채널 이상의 다채널 오디오 재생 환경을 염두에 둔 실감 음향 콘텐츠 저작 시에도 객체화 되어 있는 입력 신호는 보다 능동적이고 현장감 있는 콘텐츠 제작에 필수적인 요소임이 분명하다.

이와는 별도로, 여러 악기가 혼합되어 있는 것이 보통인 음악 신호에서 유용한 정보를 획득하기 위해서는 분리되어 있는 음원을 활용하는 경우 보다 정교한 정보 획득이 가능하다. 음악 콘텐츠가 디지털화되면서, 방대한 양의 음악 콘텐츠 중에서 사용자가 실제로 듣고 싶어 하는 음악을 보다 효율적으로 찾아주는 것에 대한 필요성이 점차로 대두되고 있으며, 보다 지능화된 서비스를 위해서 다양한 정보를 보다 효율적으로 추출하는 방법에 대한 연구가 활발히 진행되고 있다[3]. 특히 음악 음원 분리의 결과는 특정 음원에 대한 정보를 보다 독립적으로 추출할 수 있도록 함으로써 최종적으로는 다양한 목적의 음악 분류 시스템[4] 또는 예를 들어, 허밍에 의한 질의(Query by Humming) 등의 응용에서 시스템 전체의 성능을 높이는 데에 효과적으로 사용될 수 있다.

전술된 바와 같이 음악 음원 분리 결과물의 잠재적인 활용성은 다양한 분야에 걸쳐 있음에도, 음악 음원 분리는 아직까지 그 분리 성능이 여러 응용에서 요구하는 바에 못 미치는 경우가 많다. 음악 신호에 한정되지 않은 기존의 음원 분리 연구는 주로 복수의 음원이 혼합되는 여러 경우의 수에 따른 복수의 혼합 신호를 확보한 상황을 가정하며, 혼합 과정을 모델링하기 위한 행렬을 가정하고 그것을 추정하는 방식으로 연구되어 왔다. 반향이 있는 일반적인 녹음 환경에서의 음원 분리 역시 주파수 영역 변환 등을 통해 시간 영역에서의 필터 연산을 행렬 연산으로 표현하여 추정할 수 있다는 점에서, 그 방법론에 많은 다양성이 있으나, 기본적으로 음악 음원 분리 자체가 가지고 있는 독특한 제한 사항은 기존 방식을 그대로 적용하기에 부적합하다. 먼저 시간 영역 또는 주파수 영역에서의 혼합 과정 모델링을 위한 행렬 추정의 경우, 적어도 음원 개수만큼의 혼합 신호가 확보되어야 한다는 문제가 있으며, 이는 스테레오 채널을 통해 2개의 혼합 신호만을 확보할 수 있는 음악 음원 분리 환경에서의 큰 제약으로 작용한다. 또한, 반향이 없거나, 반향이 있어도 음원과 센서의 위치가 바뀌지 않고 녹음실의 구조도 바뀌지 않는 상황에서 하나의 필터로 녹음 환경을 모델링할 수 있다는 점을 주로 강조하는 기존의 음원 분리 방식의 가정에 비해, 음악 신호의 경우 음반 제작과정에서 적용되는 인공적인 필터 연산이 시간에 따라 변할 수 있다는 점 등 고착된 환경을 예측하기 어렵다는 문제점이 있다.

음악 음원 분리의 경우 이러한 제약을 바탕으로, 주로 기존의 녹음 환경 모델과는 별도로, 음원 자체가 가지고 있는 특성을 이용하는 방식에 대한 연구가 있어왔다[5~6]. 그러나 분리가 가능한 음원의 종류를 한정하지 않는 경우는, 실제 음악에 적용해서 보편적으로 좋은 성능을 낳기가 어려우며, 주로 가창 음원과 타악기 음원을 따로 추출해 내는 연구가 활발히 이루어져 왔다.

가창 음원의 경우, 두드러지는 음원의 음정을 추정하고 추정된 음정을 기반으로 해당 음원의 주파수 분포를 획득하여 마스크 또는 행렬 분해 기법을 활용하여 분리하는 방법이 발전되어 왔다[7~8]. 이 경우, 혼합 신호에서 추정하는 두드러지는 음원의 음정 추정이 부정확한 경우가 많다는 문제, 음악 신호가 스테레오 신호임에도 불구하고 이러한 채널 정보를 활용하지 않는다는 점 등의 개선의 여지가 있다. 이에 반해 스테레오 정보만을 활용하는 센터 채널 추출(Center Channel Extraction) 방식의 경우, Adobe사의 Audition 소프트웨어[9]를 포함한 다양한 상용 응용 프로그램들이 구현하여 제공하고 있는 일반적인 방법이다. 그러나 가창 신호가 스테레오 음상 범위 중 가운데에 위치한다는 가정이 언제나 일반적이지 않다는 점은 제외하더라도, 공간감의 연출을 위해 각 음원 별로 음향 효과를 주는 현대의 음악 신호는 스테레오 음상 정보만을 이용해서 완전히 분리하기에는 어려운 점이 많다.

타악기 음원은 특수한 경우를 제외하고는 넓은 주파수 대역에 분포한 잡음에 가까운 신호가 강한 타격에 의해 짧은 시간 동안 출현했다가 사라지는 고유의 특성을 보인다. 이러한 특성을 활용하여, 행렬 분해된 시간-주파수 영역 혼합 신호의 구성 성분이 타악기로부터 기반한 것인지 화성악기로부터 기반한 것인지를 판별하는 분류 기법과[10], 이미 존재하는 타악기 솔로 연주 신호를 사전정보로 활용하는 비음수 행렬 공동 분해를 이용하는 방법[11~12] 등이 연구되어 왔다.

본 논문에서는 스테레오 음상 정보를 이용하는 가창 음원의 분리 방식이 가지고 있는 근본적인 취약점을 지적하고, 이를 해소하기 위해 모노 신호를 기반으로 하는 가창 신호 분리 방식과의 연동을 제안한다. 모노 신호에서의 가창 신호 분리 방식 역시 전술된 바와 같이 그 자체로써 취약점이 있으나, 이종의 분리 방식이 서로 통합됨으로써 각각의 약점이 보완될 수 있는 새로운 구조를 제안한다.

본 논문은 다음과 같이 구성된다. Ⅱ 장에서는 기존의 가창 신호 분리 방식인 센터 채널 추출 방식과 모노 채널에서의 음정 기반 분리 방식에 대한 설명을 제공하며 각각의 문제점 또한 설명한다. Ⅲ 장에서는 본 논문에서 제안하는 통합 구조를 설명한다. Ⅳ 장에서는 실제 상용 음악 신호를 대상으로 제안하는 통합 구조가 분리 성능을 향상시킴을 입증하는 실험 결과들을 제시하며, Ⅴ 장에서는 결론 및 향후 연구 계획을 제시한다.

Ⅱ. 가창 음원의 분리를 위한 기존의 독립적 방식

1. 센터 채널 추출

센터 채널 추출 방식은 상용 스테레오 음악 신호에서 왼쪽 또는 오른쪽 채널에 보다 치우쳐진 음상을 가지는 다른 악기 음원과 코러스 음원에 비해, 중요 가창 음원은 가운데 음상에 몰려서 위치하는 경향성을 이용해서 가창 음원을 분리하는 방식이다. 분리된 이후의 음질 개선을 위해 다양한 전처리 또는 후처리 과정이 추가될 수 있으나, 기본적으로는 양 채널간의 음량 차이와 위상 차이 정보를 이용하는 방식이 가장 일반적이라고 할 수 있다. 먼저 이산 퓨리에 변환(Discrete Fourier Transform) 등의 시간-주파수 변환 과정을 통해 변환된 신호 중 특정 프레임 t와 특정 주파수 ω에서의 신호 성분을 $X(t,\omega)$라고 할 때, 왼쪽 채널과 오른쪽 채널의 차이를 이용하여 양 채널이 가지는 음량의 차이를 제한할 수 있다.

$$|\log_{10}(|X_L(t,\omega)|)-\log_{10}(|X_R(t,\omega)|)|<\lambda \qquad (1)$$

식 (1)에서 $|X_L(t,\omega)|$과 $|X_R(t,\omega)|$은 각각 왼쪽 채널과 오른쪽 채널의 시간-주파수 영역 절대값을 나타내며, λ는 미리 정해진 기준값으로, 그 값의 크고 작음에 따라 센터 채널로 간주되는 범위가 바뀐다. 식 (1)의 부등식이 성립하는 경우의 특정 프레임 t와 주파수 ω의 성분은 센터 채널에 위치한다고 간주한다.

또한 음량 정보 이외에도 위상의 차이를 이용해서, 유사한 방식의 판별을 할 수 있다.

$$|\Phi(X_L(t,\omega))-\Phi(X_R(t,\omega))|<\phi \qquad (2)$$

식 (2)에서 $\Phi(X)$는 복소수 X의 위상값을 의미하며, ϕ는 식 (1)의 λ와 마찬가지의 기준값이다.

상기와 같은 시간-주파수 영역 신호의 판별 기준은 상용 음악의 센터 채널 추출에 있어서 아래와 같은 문제점을 가질 수 있다. 먼저 주요 가창 음원과 피아노, 기타 등 3개의 음원으로 이루어진 혼합 신호의 스테레오 음상 위치를 다음과 같이 가정해 보자.

그림 1은 일반적인 음악 신호에서 주요 가창 음원이 센터 채널에 위치하는 경우에 관한 예이다. 가정에 따르면, 피아노, 기타 등 다른 악기는 중간이 아닌 왼쪽 또는 오른쪽으로 치우친 위치에 있으므로, 상기 식 (1) 또는 (2)에 의해 선별된 시간-주파수 영역 성분은 역이산 퓨리에 변환을 통해 주요 가창 음원의 신호로 복원될 수 있다. 그러나 그림 1은 각각의 음원이 공간감을 위한 효과가 적용되지 않았다는 가정에 기반한 것이고, 만일 각각의 음원이 녹음 과정 또는 믹싱, 마스터링의 과정을 거치면서 잔향 효과가 더해지는 경우 아래

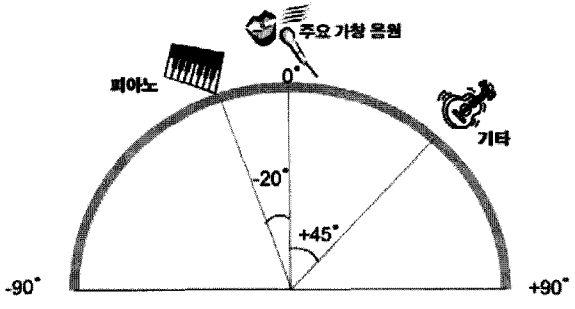

그림 1. 주요 가창 음원 및 피아노, 기타 연주로 이루어진 스테레오 공간 내에서의 음상 위치 예

Fig. 1. Example of stereophonic distribution of main vocal, piano, and guitar.

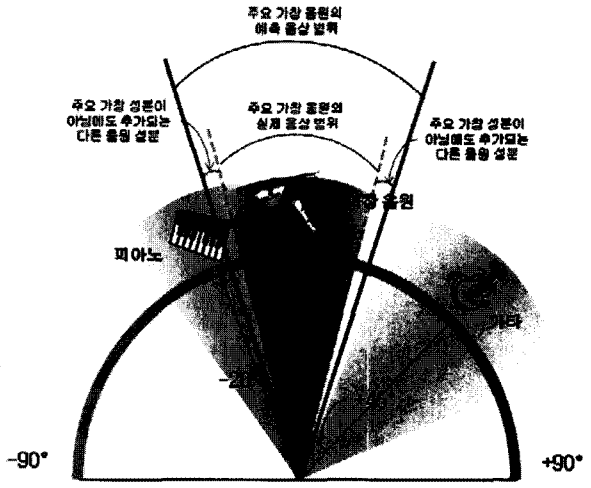

그림 2. 주요 가창 음원의 음상 범위를 실제보다 넓게 예측하는 경우 예

Fig. 2. Example of wider expectation of range of main vocal than the real one.

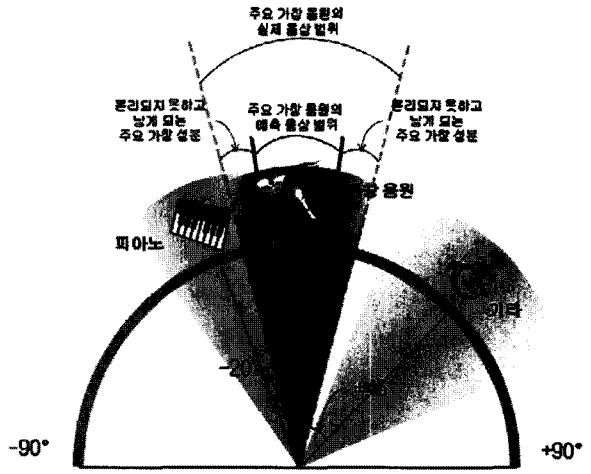

그림 3. 주요 가창 음원의 음상 범위를 실제보다 좁게 예측하는 경우 예

Fig. 3. Example of narrower expectation of range of main voccal than the real one.

그림 2-4와 같은 문제가 발생할 수 있다.

먼저, 가창 신호는 녹음시 모노 신호로 녹음되지 않을 수 있으며, 또는 모노 신호로 녹음되었더라도 음반 제작 과정에서 공간감을 위한 각종 음향 효과를 통해 정 가운데 위치를 중심으로 좌우로 약간씩의 공간감을 위한 성분 분포를 가진다. 이러한 공간감은 주로 2개의 서로 다른 필터를 원래의 모노 가창 신호에 취함으로써 야기되는 2개의 가창 신호 채널 사이의 차이에서 비롯된다고 할 수 있다. 음반 제작 과정에서 추가되는 이러한 필터 연산은 주요 가창 음원의 스테레오 공간 위치를 가운데에서 분산시킴으로써, 주요 가창 음원의 성분

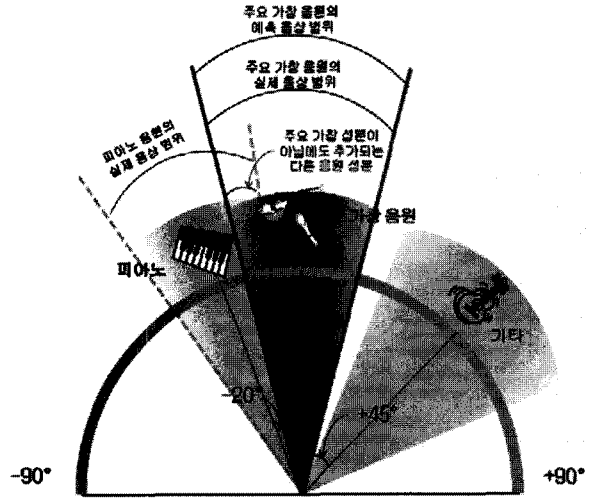

그림 4. 주요 가창 음원의 음상 범위를 정확히 예측하는 경우 예

Fig. 4. Example of exact expectation of range of main vocal.

분포를 정확히 예측하는 것을 어렵게 만든다.

그림 2는 주요 가창 음원의 스테레오 공간상 분포 범위를 실제보다 넓게 예측하는 경우이다. 이 경우, 식 (1) 또는 (2)와 같은 방식으로 센터 채널을 추출하는 경우, 주요 가창 음원 외에 인접해 있는 피아노 음원의 성분이 함께 추출될 우려가 있다. 그림 3은 이와는 반대로 주요 가창 음원의 분포를 좁게 추정하는 경우로, 이 경우 주요 가창 음원을 효과적으로 모두 분리해내지 못한다는 문제가 발생한다.

그림 4는 주요 가창 음원의 분포를 정확히 예측하는 경우에도 여전히 발생할 수 있는 문제의 예이다. 주요 가창 음원 외의 다른 악기 연주 신호 역시 스테레오 음상의 가운데에 위치하지는 않더라도 공간감을 위해 넓게 분포될 수 있으며, 그 분포 범위가 주요 가창 음원의 분포와 겹치는 경우가 생길 수 있다. 이런 경우, 주요 가창 음원의 분포를 정확히 예측했다고 하더라도 다른 악기의 음원이 함께 분리되는 그림 2와 비슷한 결과를 낳는다.

이와 같은 문제를 해소하기 위해 본 논문에서는 센터 채널 추출 시 주요 가창 성분의 범위를 좁게 추정하는 것과, 그렇게 함으로써 미처 분리되지 않는 주요 가창 음원의 성분을 추가적으로 분리하는 방법을 제안한다.

2. 음정 기반의 가창 음원 분리 방법

음정 기반의 가창 신호 분리 방식[7]은 주로 모노 신호를 대상으로 한다는 점에서 스테레오 정보를 모두 이

용하는 센터 채널 추출 방식과 차이가 있다. 음정 기반의 가창 신호 분리 방식은 먼저 혼합 신호에 포함되어 있는 다양한 음원의 음정 중 목표하는 음원(이 경우에는 가창 음원)의 음정을 프레임 별로 최대한 정확히 추정하는 것이 선행되어야 한다. 추정된 음정을 바탕으로, 음원이 유발하는 배음 위치를 쉽게 추정할 수 있으며, 배음 위치에 해당되는 주파수 성분은 목표하는 가창 음원에서 비롯된 것이라고 가정하고 마스킹(masking)한다. 마스킹된 위치에 있는 시간-주파수 성분은 기본적으로 가창 음원의 성분이라고 가정할 수 있으나, 보다 강인한 분리를 위해 마스킹하고 남은 나머지 신호를 이용하여 바이너리 가중치를 적용한 비음성 행렬 분해(BWNMF: Binary Weighted Nonnegative Matrix Factorization) 방식을 적용할 수 있다. 이 때 가중치 값은 가창 음원에서 비롯된 배음 성분을 0으로 하고, 나머지 시간-주파수 영역 성분들을 1로 하는 바이너리(binary) 마스크를 적용한다.

그림 5는 이와 같은 음정 기반의 가창 음원 분리 방식의 흐름도이다.

음정 기반의 가창 음원 분리가 성공적이기 위해서는 먼저 가창 음원의 정확한 음정 추정이 필수적이며, 성능에 큰 영향을 미친다. 음정 추정 방식은 여러 가지가 있을 수 있으나, 음악 혼합 신호에서 가창 음원의 음정을 추정하기 위해서는 크게 두 가지 정도의 난점을 극복해야 한다.

먼저 음악 신호는 여러 악기가 동시에 연주되는 구간이 대부분이기 때문에, 하나의 프레임 내에는 복수의 음정이 존재한다. 복수 개의 음정 중에서 어떤 음정이 가창 음원으로부터 비롯된 것인지를 판별하는 것은 어려운 문제이다. 각각의 악기가 음정의 위치와 관계없이 일정한 음색 정보를 가지고 있고, 이 일정한 음색 정보를 이용할 수도 있으나, 가창 음원의 경우는 일부 자음

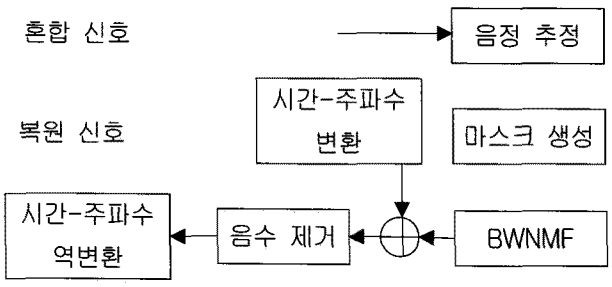

그림 5. 음정 기반의 가창 신호 분리 방식 흐름도
Fig. 5. Flow of pitch-based main vocal separation method.

에 의한 잡음성 신호와 음정을 가지고는 있으나 발음에 의해 다양하게 음색이 달라지는 유성음이 혼재되어 있는 특성을 가지고 있기 때문에 여러 음정 중 가창 음정을 판별하는 것에 문제를 야기한다. 주로 사용될 수 있는 방식은 여러 음정 중 그 음량이 두드러지는 음정을 가창 음원의 음정이라고 가정하는 방식이지만, 가창 음정이 두드러지지 않는 프레임에서는 적중률이 떨어지며, 아래 두 번째 문제와 복합적으로 또 다른 문제를 야기한다.

음악 신호에서 가창 음정을 추정할 때 생길 수 있는 두 번째 문제점은, 가창 음원이 연주되지 않는 구간에서 발생한다. 상용 음악에서는 주로 1, 2회의 간주가 삽입되는데, 이 구간에서는 가창 음원이 연주되는 대신 단독 악기가 주요 멜로디를 연주하는 경우가 많으며, 이와는 별도로 가사의 각 단어 사이 등에서 짧지만 잦게 발생하는 가창 음원의 묵음 구간은 음정 추정치에 추가적인 오류를 발생시킨다. 가창 음원이 실제로는 연주되지 않았으나, 가창 음원이 연주되는 구간을 자동화된 방식으로 미리 알고 있지 않으면, 음정 추정 알고리즘은 가창 음원이 없는 구간에서도 음정을 추정하게 되며, 이렇게 해서 추정된 음정은 가창 음원의 복원 시 묵음 구간에서 불필요한 반주 신호를 생성하게 된다.

소개된 음정 추정 기반의 가창 음원 분리 알고리즘과는 별개로, 최근 소스-필터(source-filter) 기반의 알고리즘[8] 등 모노 채널에서 가창 음원을 분리하는 알고리즘이 소개되고 있으나, 기본적인 아이디어 및 제한사항은 전술된 바에서 크게 벗어나지 않는다고 볼 수 있다.

Ⅲ. 통합 가창 음원 분리 알고리즘

본 장에서는 Ⅱ장에서 기술된 가창 음원 분리 알고리즘들의 단점을 극복하기 위한 통합적인 알고리즘을 제안한다. 통합 알고리즘은 기본적으로는 스테레오 음상 정보를 이용한 센터 채널 추출 방식의 분리를 수행하되, 그림 2~4에서 제시된 바와 같은 센터 채널 추출 방식의 문제를 해소하기 위해서 Ⅱ장 2절에서 소개된 것과 같은 음정 기반의 가창 신호 분리 방식을 개선한 다음 후처리 과정으로서 추가 적용함을 골자로 한다.

그림 6은 통합 가창 음원 분리 알고리즘의 흐름도이다. 먼저 입력된 스테레오 형식의 혼합 신호는 가운데 위치에 대부분의 가창 음원이 분포하고 있다는 가정 하

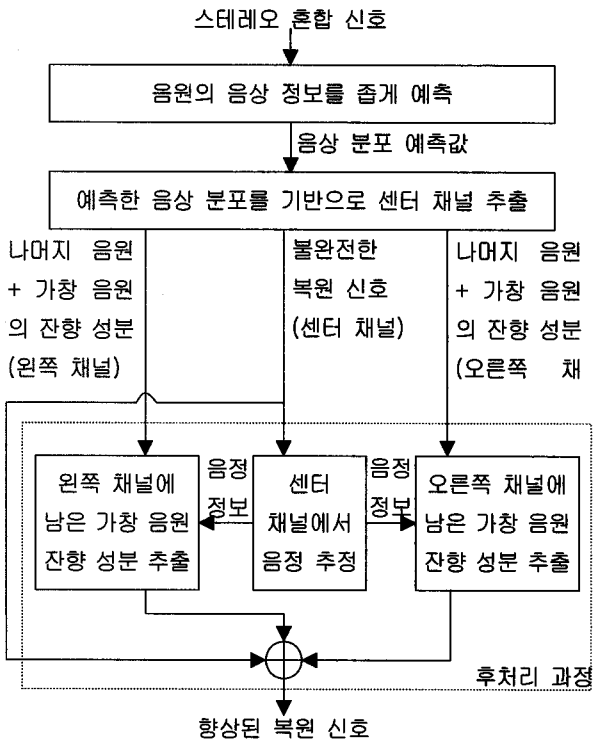

그림 6. 통합 가창 음원 분리 알고리즘 흐름도
Fig. 6. Flow of unified main vocal separation method.

에 센터 채널 추출 방식을 통해 가창 음원을 분리하게 된다. 이 때, 그림 3에서와 같이 가창 음원의 분포 범위를 좁게 가정하고, 그럼으로써 분리되고 남은 서라운드 채널 신호에 가창 음원이 남아있는 분리 환경을 조성한다. 이러한 제한은, 그림 2 또는 그림 4와 같이 센터 채널 추출 이후의 불완전한 복원 신호가 다른 음원 성분을 포함하는 것을 최소화함으로써, 이후의 음정 정보를 이용한 후처리 과정의 효율을 높이기 위함이다.

그림 6에서 점선으로 표시된 후처리 과정은 기본적으로는 Ⅱ.2 절에서 소개된 것과 같은 음정 기반의 가창 신호 분리 방식을 따른다. 그러나 후처리 과정은 센터 채널 추출 이후 나머지 신호의 왼쪽, 오른쪽 신호를 입력 신호로로 삼아 잔향 가창 음원 성분을 추가적으로 분리하는 방식으로 각각 수행된다. 모노 혼합 신호에서 가창 음원을 분리하는 기존 방식과는 달리, 음정 정보는 센터 채널 추출 결과를 추가적인 입력 신호로 삼아서 그것으로부터 얻어내고, 마스킹을 통한 음원의 복원은 각각의 서라운드 채널에 대해 따로 적용된다.

그림 7은 이 같은 과정을 좀더 자세히 도식화한 것이다. 혼합 신호를 입력으로 삼는 그림 5의 기존 방식 대비, 그림 7에서 제안하는 구조의 장점은 아래와 같다.

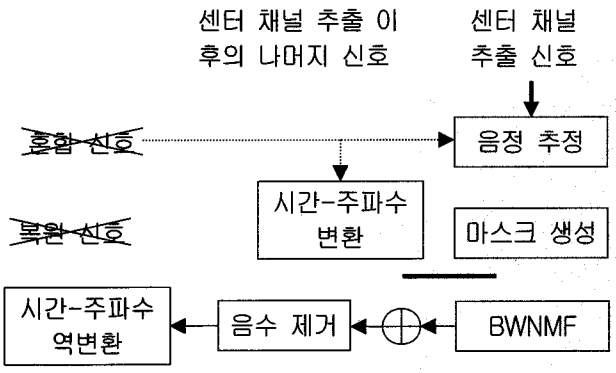

그림 7. 센터 채널 추출 이후 불완전한 복원 신호의 보강을 위한 후처리 구조. 그림 5의 음정 기반의 가창 신호 분리 알고리즘 대비 새로워진 입출력 및 흐름은 굵은 화살표로 표시. 기존 흐름과 입출력 중 중 삭제된 부분은 점선 및 X으로 표시
Fig. 7. Flow of post-processing method for reinforcing center-channel-extracted vocal source. Bold arrows represent newly inserted input, output, and flows. Unnecessary parts of previous system in fig. 5 are marked by dotted lines and Xs.

- 센터 채널 추출을 통해서 얻어낸 가창 음원의 복원 신호는, Ⅱ.1 절에서 설명된 바와 같이 불완전한 복원을 제공한다. 특히 제안되는 구조는 인위적으로 그림 3에서와 같은 상황을 야기하면서, 가창 음원이 불완전하게 분리되는 결과를 낳는다. 그러나 좁게 가정한 가창 음원의 분포 범위 덕분에, 분리된 신호는 보다 순수한 가창 음원을 포함하고 있을 확률이 높으며, 혼합 신호 대신 이 신호를 음정 추정에 활용함으로써 복수의 음정 추정에 따른 오류를 줄여 음정 추정의 정확도를 높일 수 있다.
- 그림 3과 같은 상황이 인위적으로 조성되는 제안 구조 하에서 발생하는 불완전한 복원 신호를 보강하기 위해서, 그림 7은 보다 정확히 추정된 음정 정보를 바탕으로 서라운드 채널에서 추가적인 가창 음원 분리 작업을 수행한다. 서라운드 채널은 센터 채널 추출이 이루어지고 남은 2개의 양쪽 채널이며, 이 신호에는 대부분의 반주 음원과, 공간감을 위해 분산되어 있던 가창 음원의 잔향 신호가 포함되어 있을 확률이 높다. 센터 채널 추출 결과로부터 얻은 음정 정보는 이러한 서라운드 채널 입력에 대해 바

이너리 마스크된 비음성 행렬 분해를 수행함으로써, 가창 음원의 잔향 신호를 추가적으로 분리해낸다.

Ⅳ. 실험 결과

본 장에서는 실제 음악 입력을 대상으로 제안하는 구조의 우수성을 입증한다. 입력 신호는 44.1kHz 샘플율, 16 비트로 인코딩된 일반 CD 형태의 음반과 같은 스테레오 PCM 신호이며, 상업적으로 발매된 국내 가요에서 최대한 많은 악기가 포함된 구간을 발췌하여 사용하였다. 가창 음원의 분리 성능을 측정하기 위해서 혼합 이전의 가창 음원이 확보되었으며, 채널별로 신호 대 잡음비(SNR: Signal to Noise Ratio)를 계산하여 평균을 구하였다.

$$SNR = 10\log_{10}\left(\frac{\sum_{t}(s(t))^2}{\sum_{t}(s(t)-\hat{s}(t))^2}\right) \quad (3)$$

식 (3)에서 $s(t)$는 시간 영역의 원본 가창 음원을 의미하고, $\hat{s}(t)$는 분리 과정을 통해 복원한 가창 음원을 나타낸다. 식 (3)은 양쪽 채널에 대해 각각 적용되며, 구해진 SNR 값에 대해 평균을 취한다.

SNR은 복원된 신호의 품질을 측정하기에 완벽한 방식은 아니다. 음원 분리 과정에서 고려되어야 할 각종 오차 신호를 고려한 보다 정교한 복원 신호의 모델은 식 (4)와 같이 표현할 수 있다.

$$\hat{s} = s_{\text{target}} + e_{\bigcap \text{f}} + e_{\text{noise}} + e_{\text{artif}} \quad (4)$$

식 (4)에서 $\hat{s}$는 복원된 음원, s_{target}는 목표 음원을 나타내며, e_{interf}, e_{noise}, e_{artif}는 각각 목표 음원이 아닌 다른 음원의 성분, 간섭 잡음, 분리 알고리즘에 의해 발생하는 음악적 잡음(musical noise)와 같은 잡음을 의미한다[13]. 따라서, 원본과의 비교 대상이 되는 간섭 신호의 종류에 따라, 신호 대 왜곡비(Signal to Distortion Ratio), 신호 대 간섭 음원비(Signal to Interference Ratio) 등 다양한 측정 방법이 제안되어 있다.

반면, 대부분의 음원 분리 문제에 있어서, 복원 신호에 포함된 각종 간섭 신호를 상기 식 (4)와 같은 모델로 구분하는 것은 어려우며, 경우에 따라서는 세 가지 간섭 신호 중 일부만이 포함되기도 하는 등의 다양한 가능성이 존재한다. 본 논문에서는 음악 음원 분리 알고리즘의 성능 평가에 일반적으로 활용되고 있는 방식[7, 10~12]의 일환으로, 목표 음원과 복원 신호의 차이를 각종 간섭 신호의 합이라고 간주하고, 이를 SNR이라고 명명하는 평가 방식을 따랐다.

Ⅳ장 1절의 실험에는 10초짜리 신호가 사용되었으며, Ⅳ장 2절의 실험에는 10곡의 다른 노래에서 발췌한 10초길이의 구간 10개가 사용되었다.

1. 다양한 음상 분포 예측값에 따른 성능 변화

본 절에는 주어진 한 가지의 신호에 대해, 센터 채널 추출 파라미터를 다양하게 변경하면서 그에 따른 제안 구조의 성능 변화를 살펴본 실험 결과를 제공한다.

센터 채널 추출을 위해서 식 (1)과 (2)에서 설명된 바와 같이 음량에 의한 판별 기준과 위상에 의한 판별 기준값 λ와 ϕ를 각각 1 ~ 10dB, 1 ~ 61°까지 변화시키면서 각각의 변화에 따른 SNR 값의 변화 추이를 조사하였다.

본 절에서는 원활한 파라미터 변경을 위해서, 성능이 좋은 상용 제품 대신 식 (1)과 (2)의 기준만으로 자체 구현한 센터 채널 추출 알고리즘을 적용하였다.

10초 길이의 입력 신호는 먼저 구현된 센터 채널 추출 방식을 통해 센터 채널과 서라운드 채널로 분리되며, 이 때 센터 채널 추출은 각각의 파라미터 조합에 따라 조금씩 다른 분리 결과를 낳는다. 파라미터 조합 별 센터 채널 분리 결과는 전술된 통합 구조에 의해 후처리 단계가 추가적으로 적용되며, 이 때 센터 채널 신호에서 추출된 음정 정보를 이용하여 서라운드 채널 신호에 남아 있는 가창 음원의 잔향을 추가적으로 분리한 다음 센터 채널 신호에 합침으로써 센터 채널 신호를 개선시킨다.

그림 8과 9는 파라미터 조합 별로 획득한 복원 신호의 SNR 값을 시각화한 것이다. 그림 8은 센터 채널 추출만을 적용한 경우의 파라미터 조합 별 SNR 값을 나타내며, 그림 9는 추가적으로 음정 기반 후처리 모듈을 적용한 결과이다. 두 그림에서 모두 가로 축은 위상 판별 기준 ω의 변화를 나타내며, 세로 축은 음량 판별 기준 λ의 변화를 나타낸다. 또한 색이 진해질수록 더 높은 SNR 값임을 의미한다.

그림 8과 9는 전반적으로 볼 때 단순한 센터 채널 추출 보다는 본 논문에서 제시하는 통합 구조가 보다 우수하다는 것을 입증한다. 즉, 다양한 범위의 센터 채널 추출 파라미터 세팅에 의해 센터 채널의 판별 기준을

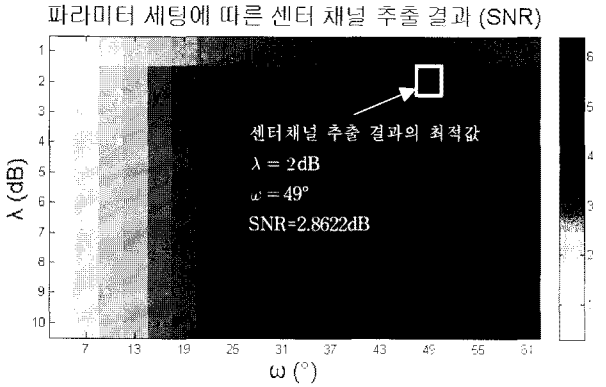

그림 8. 센터 채널 추출 신호를 가창 음원이라고 가정한 경우의 분리 성능. 음량 및 위상 파라미터 변화에 따른 분리 성능의 변화

Fig. 8. Variation of separation performance of the center-channel-extracted signal as a reconstruction of main vocal source by changing amplitude and phase discriminant parameters.

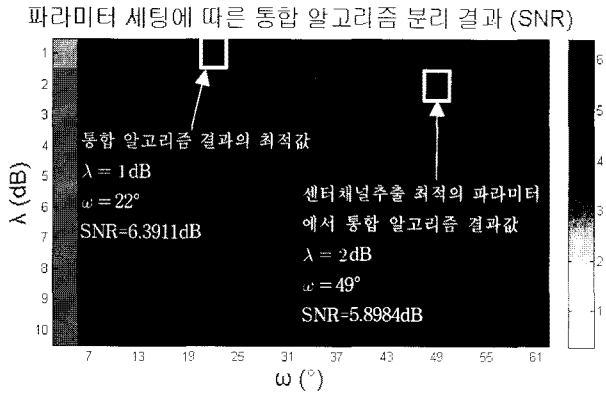

그림 9. 제안하는 통합 구조를 적용한 결과의 분리 성능. 음량 및 위상 파라미터 변화에 따른 분리 성능의 변화

Fig. 9. Variation of separation performance of the result of the proposed unified system.

다양화했지만, 그림 9의 SNR 값이 그림 8의 SNR 값보다 전반적으로 높다(색이 진하다)는 것은 제안 구조가 특정한 상황에서만 우수한 것이 아니라 다양한 분리 환경에 강인하게 적용될 수 있다는 점을 반증한다.

보다 구체적으로는, 그림 8에서 가장 높은 SNR 값인 2.8622dB를 도출시켜주는 파라미터 세팅은 λ=2dB, ω=49°일 때인데, 같은 파라미터 세팅에 대해서, 제안하는 구조를 적용하면 5.8984dB로 오히려 3.0362dB가 상승하는 결과를 보여준다. 이 같은 결과는, 실제로 센터 채널 추출을 수행할 때는 가창 음원 복원을 위한 최적의 파라미터를 예측하기 어렵지만, 최적의 파라미터를 알고 있고, 그것을 적용하는 최적의 결과라고 하더라도 본 논문에서 제안하는 통합 구조가 보다 더 우수하다는 것을 입증한다.

또한 센터 채널 추출만을 했을 때, 최적의 결과를 도출하는 파라미터 세팅은 λ=2dB, ω=49°인데 반해, 음정 추정 기반의 후처리를 적용하는 것을 감안한 최적 결과는 6.3911dB, 파라미터 세팅은 λ=1dB, ω=22°로, 서로 다른 세팅 값에서 최적의 값을 도출하며, 좁게 가정된 센터 채널 분포에서 3.5289dB의 상승분을 보인다는 것을 그림 9를 통해 확인할 수 있다. 이는 Ⅲ장에서 제안한대로, 센터 채널 추출 시 가창 음원의 분포 범위를 조금 좁게 가능하고, 서라운드 채널에 남아 있는 가창 음원의 잔향 성분을 추가적으로 추출하는 것이 보다 높은 성능을 보인다는 것을 의미한다.

2. 가창 음원의 분포를 좁게 고정시킨 경우

본 절에서는 Ⅳ장 1절에서의 실험 결과를 바탕으로, 센터 채널의 범위를 좁게 가정하여 추출한 다음, 그 결과에 후처리 과정을 적용한 실험 결과를 제시한다. 다만, 본 절에서의 실험은, 10 개의 다양한 음악 콘텐츠에 대해서 수행함으로써, 제안 구조가 입력 신호의 변화에 대해 강인하다는 것을 보이고자 한다. 또한, 본 절에서의 실험에는 상용 제품인 Adobe사의 Audition 3.0 프로그램[9]을 사용하여 센터 채널 추출을 수행하였다. 그 이유는 센터 채널 추출만을 수행한 것과 그것에 추가적인 작업을 덧붙인 결과를 비교하는 만큼, 센터 채널 추출 모듈은 가능한 최상의 성능을 발휘하는 모듈을 사용하는 것이 공정한 실험이라는 판단에서이다.

Audition 프로그램에 포함된 센터 채널 추출 기능은 채널 간 음량 차이와 위상 차이를 기본적인 기준으로 하여 추출 후의 음질 향상을 위한 다양한 옵션을 제공하지만, 상용 제품에 의한 특성상 그 내부적인 알고리즘 및 작동 방식은 알려지지 않고 있다. 본 논문에서는 잘 알려져 있는 기본적인 파라미터에 대한 변경을 통해 주장하는 분리 환경을 조성하고자 한다. 본 절에서의 실험은 아래와 같은 5가지 단계로 수행되었고, 각 단계의 결과가 원본 가창 음원과 비교되었다.

1. 혼합 신호를 가창 음원의 복원 신호로 가정하고 구한 SNR (음원 분리를 수행하지 않은 경우로, 분리 수행을 통한 성능 향상폭 측정을 위함)
2. Audition의 기본 파라미터 세팅(λ=1dB, ω=5°)을 통한 센터 채널 추출
3. Audition의 기본 파라미터 세팅을 통한 센터 채널

추출(2번 단계) 이후 음정 기반 후처리 모듈 추가 수행

4. 기본 세팅보다 좁게 바꾼 파라미터 세팅(λ=0.5dB, ω=1.5°)을 통한 센터 채널 추출
5. 기본 세팅보다 좁게 바꾼 파라미터 세팅을 통한 센터 채널 추출(4번 단계) 이후 음정 기반 후처리 모듈 추가 수행

표 1은 상기 실험 결과를 나열한 것이다. 먼저 첫 번째 열은 음원분리가 수행되지 않은 경우, 즉 혼합 신호를 복원 신호로 가정한 경우의 SNR 값이다. 평균적으로 음의 SNR 값(-1.0420dB)을 가지는 이 경우는, SNR 계산식에서 분모의 값이 분자의 값 보다 더 큰 경우로, 원본 가창 음원에서 혼합 신호를 뺀 나머지 신호에 다른 악기 신호가 상당히 많이 포함되어 있기 때문에 발생하는 현상이다.

Audition의 기본 파라미터 세팅을 통해 추출한 센터 채널을 가창 음원의 복원 신호로 가정한 두 번째 열은, 첫 번째 열에 비해 평균 3.0729dB가 향상된 결과를 보여주는데, 이는 센터 채널 추출 방식이 기본적으로 가창 음원의 복원에 어느 정도의 성능 향상을 가져온다는 것을 알 수 있다.

세 번째 열은, Audition의 기본 파라미터 세팅을 통해 추출한 센터 채널과 서라운드 채널을 입력 신호로 삼아 본 논문에서 제안하는 통합 구조를 적용한 결과이다. 두 번째 열의 결과에 비해 오히려 0.0518dB 성능이 떨어진 이 결과는, Audition의 기본 파라미터 세팅이 제안 알고리즘을 적용하기에 부적합할 정도로 넓게 센터 채널 분포를 가정한다는 사실을 입증한다.

네 번째 열은, 기본 파라미터 세팅보다 더 좁게 설정된 센터 채널 추출 결과물의 SNR 값이다. 좁게 설정된 센터 채널 추출 결과물은 보다 순수한 가창 음원을 포함하고 있으나, 2~3열에 비해 각각 평균 0.1452, 0.1970dB가 떨어진 결과를 낳았다. 이는 서라운드 채널에 가창 음원의 잔향 신호가 분리되지 않고 많이 남아 있다는 것을 의미하고, 그림 3의 상황과 부합한다고 볼 수 있다.

다섯 번째 열은 좁게 설정된 센터 채널 추출 결과물과 해당 서라운드 채널 신호를 입력으로 삼아 제안하는 후처리 모듈을 적용한 최종적인 결과이다. 이 결과는 평균적으로 앞선 실험 결과들 모두를 월등히 능가하는 것으로, 제안되는 구조의 우수성을 입증하는 동시에, 제안되는 구조가 올바로 작동하기 위해서는 센터 채널 분포를 인위적으로 좁게 조절하는 것이 좋다는 가설을 증명한다. 개별적인 입력 신호별로 비교해 보아도, 5, 7 번 곡을 제외하면 모든 노래에서 성능이 앞선다는 사실을 알 수 있다 (굵은 글씨체로 표시).

그림 10, 11, 12는 각각 6번 곡에 대해 기본 파라미터 세팅값으로 센터 채널 추출만을 적용한 결과 (표 1의

표 1. 10 곡의 상용 음악 신호 입력에 대해 제안하는 통합 구조를 적용한 결과의 분리 성능
Table 1. Separation performance of the proposed unified system for 10 commercial music mixtures.

노래	혼합 신호 (dB)	센터 채널 추출 (기본 파라미터 λ=1dB, ω=5°) (dB)	센터 채널 추출 (기본 파라미터 λ=1dB, ω=5°) + 음정 기반 후처리 (dB)	센터 채널 추출 (좁은 파라미터 λ=0.5dB, ω=1.5°) (dB)	센터 채널 추출 (좁은 파라미터 λ=0.5dB, ω=1.5°) + 음정 기반 후처리 (dB)
1	-0.0529	3.9873	3.8071	4.0085	4.6057
2	-1.4235	2.3249	1.8516	2.6549	3.0773
3	-0.0913	1.8378	1.8438	2.3436	3.0782
4	-4.6789	-3.7655	-3.7941	-3.5971	-3.4753
5	2.2555	6.5204	6.4761	4.9527	6.0436
6	2.2282	4.0471	5.0073	2.2949	5.7894
7	-3.3109	1.6906	0.9140	2.4563	2.1276
8	-4.7293	-3.9324	-3.9142	-3.6757	-3.3961
9	1.8655	4.9130	4.9467	4.4802	6.0274
10	-2.4823	2.6858	2.6525	2.4205	4.4000
평균	-1.0420	2.0309	1.9791	1.8339	2.8278

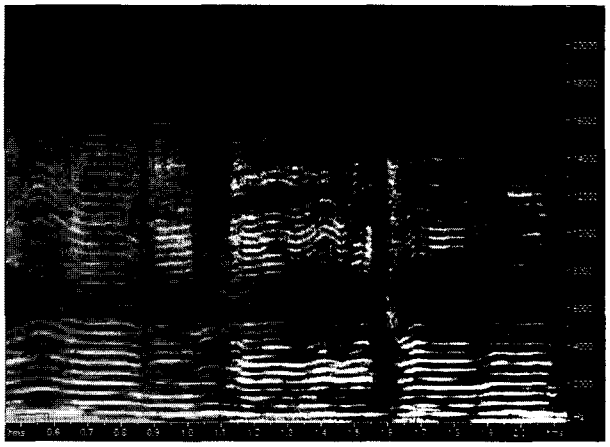

그림 10. 기본 파라미터 세팅값으로 센터 채널 추출만을 한 결과 스펙트로그램 (6번 노래, 표 1의 두 번째 열)

Fig. 10. Spectrogram of the result of the center channel extraction with the default parameter set (song number 6, column 2 in table 1).

그림 11. 제안하는 통합 구조를 적용한 결과 스펙트로그램 (6번 노래, 표 1의 다섯 번째 열)

Fig. 11. Spectrogram of the result of the proposed system (song number 6, column 5 in table 1).

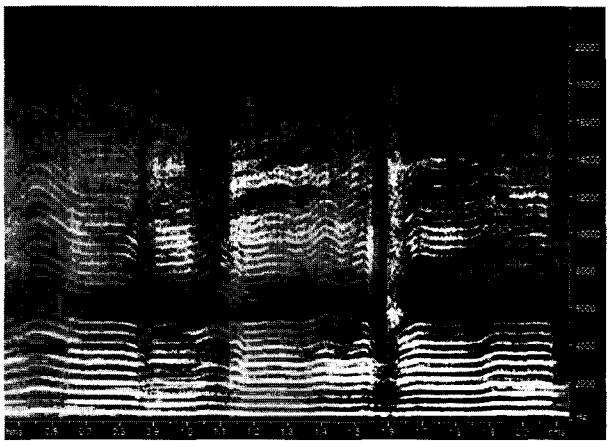

그림 12. 원본 가창 음원 (6번 노래)

Fig. 12. Spectrogram of the original vocal source (song number 6).

두 번째 열), 제안 알고리즘의 적용 결과 (표 1의 다섯 번째 열), 그리고 원본 가창 신호의 스펙트로그램이다. 스펙트로그램 상에서 확인할 수 있는 사항은, 먼저 제안된 알고리즘을 통해 고주파 대역에 주로 분포하는 음향 효과에 의한 잔향 신호가 제안 알고리즘을 통해 조금 더 수집되었다는 것이다. 또한, 특히 6kHz 이하의 저주파 대역에서 원본 가창 신호의 배음 성분이 끊어지거나, 간섭 음원의 일부 성분이 돌출되면서 발생하는 음악적 잡음 신호가 현저히 줄어들었다는 것 또한 확인할 수 있다.

참고로 본 장의 모든 실험에 쓰인 혼합 신호는 타악기 신호를 제외한 신호임을 밝혀 둔다. 실제로 타악기 신호까지 포함된 혼합 신호는 센터 채널 추출 과정에서 타악기 음원의 많은 성분이 센터 채널에 남아있게 되므로 가창 음원 분리 결과를 현저히 떨어뜨리는 문제가 있다. 향후 적절한 타악기 분리 알고리즘을 적용함으로써 개선을 기대할 수 있는 부분이다.

V. 결 론

본 논문에서는 상용 음악 신호에서 유효하게 음원 분리를 수행하기 위한 통합적인 알고리즘을 제시하였다. 음악 음원 분리에서 가장 중요하게 취급되는 가창 음원의 경우, 스테레오 음상 정보를 주로 활용하는 방식과 모노 신호에서의 음정 추정 기반 방식의 기술이 개발되어 왔으나, 각각의 방식은 장단점이 공존하는 불완전한 부분이 있었다. 본 논문에서는 기존의 독자적인 가창 음원 분리 기술을 수정 및 통합함으로써, 각각의 기술이 가진 단점을 상쇄하고 보다 높은 품질의 복원 신호를 도출하는 새로운 구조를 제시하였다. 제안된 기술은 상용 음악 입력 신호를 이용한 실험을 통해 그 우수성이 검증되었다.

향후 음악 음원 분리를 위해 보다 다양한 기술들의 특성을 파악하고, 이를 개선시키는 노력과 더불어서, 각 기술의 단점이 통합을 통해 상쇄될 수 있음을 보여주는 추가적인 연구를 수행할 예정이며, 특히 가창 음원만이 아니라 보다 다양한 음원을 대상으로 분리 알고리즘 적용 범위를 확대할 예정이다.

참 고 문 헌

[1] I. Jang, J. Seo and K. Kang, "Design of a File

Format For Interactive Music Service," ETRI Journal Letter (submitted for publication)
[2] ISO/IEC JTC 1/SC29/WG11 w11158, Text of ISO/IEC FDIS 23000-12 Interactive Music AF, MPEG, Feb. 2010.
[3] P. Chordia and A. Rae, "Using Source Separation to Improve Tempo Detection," Proc. ISMIR 2009, pp. 183-188.
[4] E. Tsunoo, T. Akase, N. Ono, and S. Sagayama, "Music Mood Classification by Rhythm and Bass-line Unit Pattern Analysis," Proc. ICASSP 2010, pp. 265-268
[5] M. Kim and S. Choi, "On spectral basis selection for single channel polyphonic music separation," in Proceedings of the International Conference on Artificial Neural Networks (ICANN), vol. 2. Warsaw, Poland: Springer, 2005, pp. 157 - 162.
[6] D. FitzGerald, M. Cranitch, and E. Coyle, "Shifted nonnegative matrix factorisation for sound source separation," in IEEE Workshop on Statistical Signal Processing, Bordeaux, France, 2005.
[7] T. Virtanen, A. Mesaros, and M. Ryynänen, "Combining Pitch-Based Inference and Non-Negative Spectrogram Factorization in Separating Vocals from Polyphonic Music," Proc. SAPA 2008.
[8] J.-L. Durrieu, A. Ozerov, C. Févotte, G. Richard and B. David, "Main instrument separation from stereophonic audio signals using a source/filter model," EUSIPCO 2009.
[9] http://www.adobe.com/products/audition
[10] M. Helén, T. Virtanen, "Separation of Drums From Polyphonic Music Using Non-negative Matrix Factorization and Support Vector Machine," Proc. 13th European Signal Processing Conference, 2005.
[11] J. Yoo, M. Kim, K. Kang, and S. Choi, "Nonnegative Matrix Partial Co-Factorization for Drum Source Separation," Proc. ICASSP 2010.
[12] M. Kim, J. Yoo, K. Kang, and S. Choi, "Blind Rhythmic Source Separation: Nonnegativity and Repeatability," Proc. ICASSP 2010.
[13] E. Vincent, R. Gribonval, and C. Févotte, "Performance Measurement in Blind Audio Source Separation," IEEE Trans. Audio, Speech, and Language Processing, Vol. 14, no. 4, pp. 1462-1469, July 2006.

저 자 소 개

김 민 제(정회원)
2004년 아주대학교 정보 및 컴퓨터공학부 학사 졸업.
2006년 포항공과대학교 컴퓨터공학과 석사 졸업.
2006년~현재 한국전자통신 연구원 연구원
<주관심분야 : 기계 학습, 음원 분리, 음악 신호 처리, 음성 및 오디오 코덱>

장 인 선(정회원)
2001년 충북대학교 전기전자 공학부 학사 졸업.
2004년 포항공과대학교 컴퓨터공학과 석사 졸업.
2004년~현재 한국전자통신 연구원 선임연구원
<주관심분야 : 객체기반 오디오 시스템, 음원분리, 오디오 부호화, 3차원 오디오 및 음향 신호처리>

강 경 옥(정회원)-교신저자
1985년 부산대학교 물리학과 학사 졸업.
1988년 부산대학교 물리학과 석사 졸업.
2004년 한국항공대학교 전자공학과 박사 졸업
2006년 영국 University of Southampton 방문연구원
1991년~현재 한국전자통신연구원 책임연구원 (실감음향연구팀장)
<주관심분야 : 오디오 신호처리, 객체기반 오디오, 3D 오디오, 음성 및 오디오 코덱>