한국음향학회지 제30권 제6호 pp. 353~359 (2011)

보컬 피치 검출의 성능 향상을 위한 보컬 강화 기술

Vocal Enhancement for Improving the Performance of Vocal Pitch Detection

이 세 원, 송 재 종*, 이 석 필*, 박 호 종
(Sewon Lee, Chai-Jong Song*, Seok-Pil Lee*, Hochong Park)

광운대학교 전자공학과, *전자부품연구원
(접수일자: 2011년 4월 29일; 채택일자: 2011년 7월 29일)

본 논문에서는 다성 음악 신호의 보컬 피치 검출 성능을 향상시키기 위해 음악 신호의 보컬 신호를 강화시키는 전처리 기술을 제안한다. 제안한 보컬 강화 기술은 입력된 다성 음악 신호로부터 반주 신호를 예측하고, 예측된 반주 신호를 입력된 보컬 신호의 크기에 맞춰 가공하여 반주 복사본 신호를 생성한다. 마지막으로 주파수 영역에서 반주 복사본 신호를 원래 다성 음악 신호에서 제거하여 보컬이 강화된 출력 신호를 생성한다. 원 음악 신호와 제안한 방법으로 보컬이 강화된 신호에 동일한 보컬 피치 검출 방법을 각각 적용하여 피치 검출의 정확도를 측정하였고, 제안한 기술에 의하여 피치 검출 정확도가 평균 7.1 % 포인트 향상된 것을 확인하였다.
핵심용어: 다성 음악 신호, 보컬 강화, 보컬 피치, 반주 제거
투고분야: 음악음향 및 음향심리 분야 (8.6)

This paper proposes a vocal enhancement technique for improving the performance of vocal pitch detection in polyphonic music signal. The proposed vocal enhancement technique predicts an accompaniment signal from the input signal and generates an accompaniment replica signal according to the vocal power. Then, it removes the accompaniment replica signal from the input signal, resulting in a vocal-enhanced signal. The performance of the proposed method was measured by applying the same vocal pitch extraction method to the original and the vocal-enhanced signal, and the vocal pitch detection accuracy was increased by 7.1 % point in average.
Keywords: Polyphonic music, Vocal enhancement, Vocal pitch, Accompaniment removal
ASK subject classification: Musical Acoustics and Psychoacoustics (8.6)

I. 서 론

최근 다성 음악 (polyphonic music) 신호에서 음악의 특성을 검출하는 기술이 다양한 분야에서 연구되고 있다. 특히 다성 음악 신호에 포함되어 있는 보컬 (vocal) 신호를 검출하는 기술은 대중음악의 DB 검색 및 카테고리 별 분류, 인식 등과 같은 응용 분야에서 핵심 기술로 사용되고 있다. 일반적으로 DB에서 음악을 검색할 때 다성 음악 신호의 주요 멜로디 정보를 사용하며, 주요 멜로디는 주로 보컬 신호가 맡기 때문에 보컬 신호의 피치 정보를 정확히 검출하는 것은 매우 중요하다. 그러나 보컬 신호와 반주 신호가 복잡한 형태로 혼합되어 있는 대중음악 신호에서 보컬 신호만을 정확히 검출하는 것은 매우 어려운 작업이다.

기존의 보컬 신호 검출 기술들은 일반적으로 다성 음악 신호를 구성하는 여러 소리 신호들의 기본 주파수 (fundamental frequency)들을 먼저 찾고, HMM (Hidden Markov Model) 또는 GMM (Gaussian Mixture Model)과 같은 수학적 모델링을 이용하여 보컬 신호를 검출하는 방법을 사용하였다 [1-4]. 특히 보컬 신호의 특성을 고려하여 기존의 수학적 확률 모델이 아닌 인간의 청각 모델 (auditory model)을 이용한 연구는 향상된 보컬 신호 검출 성능을 보여주기도 하였다 [5]. 이 기술들은 다성 음악 신호에 섞여있는 모든 소리 신호들의 피치 주파수들을 프레임 단위로 검출하고, 모델링을 이용하여 같은 악기의 피치 주파수끼리 연결 (tracking)하여 악기 신호 (instrument signal)를 생성한다. 그리고 이렇게 완성된 여러 개의 악

책임저자: 박 호 종 (hcpark@kw.ac.kr)
서울시 노원구 월계동 447-1 광운대학교 전자공학과
(전화: 02-940-5104; 팩스: 02-913-9057)

기 신호 중에서 사람의 음성 특성과 가장 유사한 신호를 찾아 보컬 신호로 최종 결정한다. 모델링을 사용하는 보컬 신호 검출 기술들은 피치 주파수를 연결하는 과정에서 발생하는 오차에 매우 취약한 문제점을 갖는다. 즉 각 프레임 마다 찾은 피치 주파수들을 정확히 모델링하지 못하면 다른 악기의 신호들이 하나의 악기 신호로 연결되고, 특히 반주 성분이 강한 영역에서는 음성 특성이 뚜렷이 나타나지 않아 보컬 신호를 검출하는데 많은 오류가 발생한다.

최근에는 모델링 기법을 사용하지 않으면서 보컬 신호를 정확히 검출할 수 있는 다양한 연구들이 발표되고 있다. 특히 다성 음악 신호의 보컬과 반주 (accompaniment) 신호 고유의 특성을 고려하여, 보컬 신호 검출 시 방해가 되는 성분들을 미리 제거해줌으로써 원하는 신호만을 쉽게 검출하게 도와주는 전처리 기술들이 개발되었다. 대표적인 전처리 기술로는 다성 음악 신호를 하모닉 부분과 퍼커시브 (percussive) 부분으로 분리하고, 보컬 신호 검출에 방해가 되는 악기의 타격 성분들을 보컬 신호 검출 전에 미리 제거해주는 방법이 있다 [6]. 그러나 이 전처리 기술은 보컬 신호 검출이라는 측면에서 여전히 문제점을 가진다. 비록 다성 음악 신호에서 퍼커시브 부분을 미리 제거해줌으로써 보컬 신호 검출을 방해하는 요소들을 제거하여 검출의 정확도 향상에 도움을 주지만 여전히 남아 있는 악기들의 하모닉 성분들은 보컬 신호 검출의 큰 걸림돌이 되기 때문이다.

따라서 본 논문에서는 다성 음악 신호를 보컬 신호와 음성 이외의 모든 악기들의 소리가 하나로 결합된 반주 신호로 이분화하고, 입력 신호를 분석하여 반주 신호를 예측하여 반주 복사본 (replica) 신호를 생성하고, 이를 입력 다성 음악 신호에서 제거함으로써 보컬 신호를 정확히 검출할 수 있게 하는 보컬 강화 기술을 제안한다. 제안한 보컬 강화 기술은 입력된 다성 음악 신호의 다섯 프레임 중에서 반주 신호의 크기가 가장 큰 프레임을 선택하고, 미리 결정된 방법을 통해 반주 복사본 신호를 생성한다. 이렇게 결정된 반주 복사본 신호를 원래 다성 음악 신호에서 제거해 보컬이 강화된 신호를 얻는다. 제안한 기술에 의하여 보컬이 강화된 신호에 기존의 보컬 피치 검출 방법을 적용하여 보컬 피치 검출의 정확도를 측정하였으며, 제안한 기술에 의하여 검출의 정확도가 평균 7.1 % 포인트 향상된 것으로 확인하였다. 즉, 기존의 보컬 피치 검출 방법을 그대로 사용하고 전처리 기술만을 적용하여 뚜렷한 성능 향상을 얻을 수 있는 것을 검증하였다.

II. 제안한 보컬 강화 기술

2.1. 개요

제안한 보컬 강화 기술에서 입력 신호는 샘플링 주파수가 8 kHz이고 모노 (mono) 신호로 가정하고, 만일 다른 규격의 신호가 입력되면 전처리 단계에서 규격을 변환시켜 입력한다. 제안한 기술의 기본 동작은 10 ms 프레임 단위로 이루어지고, 반주 신호 분석을 위하여 다섯 프레임을 하나의 슈퍼 프레임 (super frame)으로 묶어 처리한다. 제안한 보컬 강화 기술은 실시간 처리를 위한 것이 아니므로 다섯 프레임에 의한 동작 지연은 문제가 되지 않는다. 그림 1이 제안한 보컬 강화 기술의 동작 프레임 구조를 보여준다. 처음 4개의 프레임은 반주 예측의 초기화에 사용되고, 이후 5번째 프레임부터 보컬 강화가 이루어진다. 이 때 모든 처리는 다섯 개의 프레임이 결합된 슈퍼 프레임 단위로 처리한다. 즉 그림 1에서 첫 번째 보컬 강화 과정이 완료되면 슈퍼 프레임 #0에 속한 다섯 개 프레임의 보컬 신호가 모두 강화되고, 다음 두 번째 보컬 강화 과정에서는 슈퍼 프레임 #1의 다섯 개 프레임에 대한 강화 과정이 진행된다.

그림 2는 제안한 보컬 강화 기술의 전체 구조를 보여

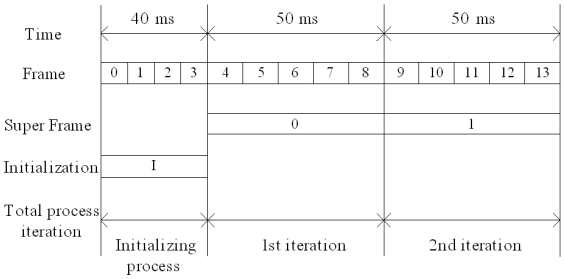

그림 1. 제안한 방법의 동작 프레임 구조
Fig. 1. Frame structure of the proposed method.

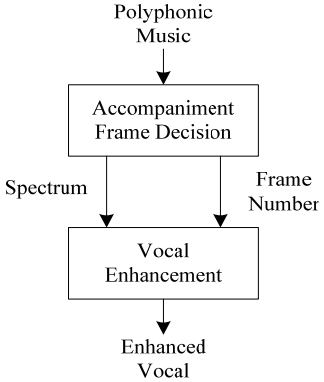

그림 2. 제안한 보컬 강화 기술의 전체 구조
Fig. 2. Overall structure of the proposed vocal enhancement method.

준다. 첫 번째 반주 프레임 결정 (accompaniment frame decision)에서는 잡음 억제 알고리즘을 이용하여 입력된 슈퍼 프레임 중에서 가장 반주 성분이 강한 프레임 한 개를 선택한다. 두 번째 보컬 강화 (vocal enhancement)에서는 이전 과정에서 선택된 반주 프레임을 이용하여 반주 복사본 신호를 생성하고, 이것을 원 다성 음악 신호에서 제거하여 보컬 신호를 강화한다. 이 때 반주 복사본 신호는 보컬 강화 과정이 반복될 때마다 슈퍼 프레임 단위로 갱신된다. 제안한 보컬 강화 기술에서 반주 신호를 예측하고 이를 제거하는 방법은 음성 신호에서의 잡음 제거에 가장 많이 사용되는 스펙트럼 감산법 (spectral subtraction)을 이용하였다 [7-8].

2.2. 반주 프레임 결정

제안한 보컬 강화 기술의 첫 번째 과정인 반주 프레임 결정은 그림 3과 같은 순서로 동작이 진행된다.

반주 프레임 결정 과정에서는 슈퍼 프레임에서 반주 신호의 크기가 가장 큰 프레임 한 개를 선택한다. 선택된 반주 프레임은 이후 반주 복사본 신호 생성에 사용된다. 각 세부 동작의 자세한 설명은 다음과 같다.

1) 전처리 동작에서는 입력된 다성 음악 신호의 샘플링 주파수를 8 kHz로 변환하고, 입력 신호가 스테레오일 경우에는 모노로 변환하고, DFT (discrete Fourier transform)를 이용하여 입력 신호를 10 ms 프레임 단위로 스펙트럼으로 변환한다. 대부분의 보컬 성분은 4 kHz 이하의 저대역에 집중되므로 일반적인 보컬 피치 검출 방법은 다운 샘플링을 통하여 불필요한 고대역을 제거하는 단계를 거친다. 따라서 4 kHz 이하 대역에서의 보컬 강화만 필요하고, 그에 따라 제안한 방법에서도 다운 샘플링 단계를 사용한다.

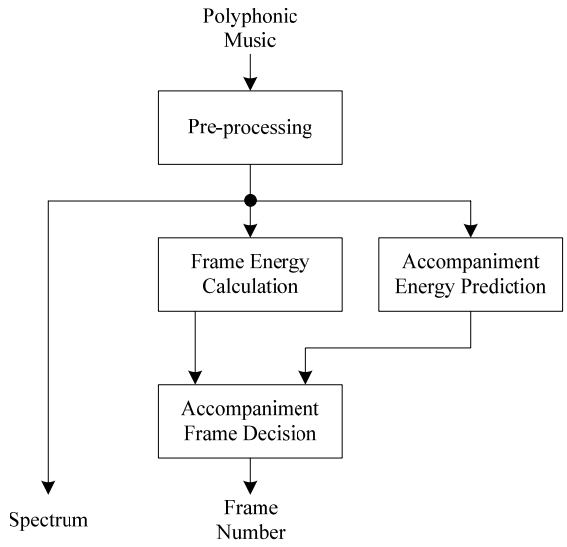

그림 3. 반주 프레임 결정의 흐름도
Fig. 3. Flowchart of accompaniment frame decision.

2) 프레임 에너지 계산 동작에서는 16개의 주파수 밴드 단위로 에너지를 계산하고, 대역별 가중치를 적용하여 모두 더하여 프레임의 전체 에너지 $FE(n)$를 구한다. n은 프레임 인덱스를 의미한다. 이 때 16개의 주파수 밴드의 대역폭과 가중치는 심리음향을 고려하여 잡음 제거에서 사용하는 음성 메트릭을 사용하여 결정한다 [7].

3) 반주 에너지 예측 동작에서는 입력 신호로부터 반주 에너지를 예측한다. 이 때 반주 에너지는 슈퍼 프레임 단위로 결정된다. 첫 번째 보컬 강화 과정에서는 그림 1과 같이 초기 4개 프레임의 평균값을 반주 신호로 결정하여 처리하고, 두 번째 보컬 강화 과정부터는 이전 보컬 강화 과정에서 결정된 AF (accompaniment flag) 파라미터를 이용하여 식 (1)과 같이 반주 신호의 에너지 $AE(m)$을 예측한다.

$$AE(m) = \begin{cases} AE(m-1) & \text{if} \quad AF=0 \\ \alpha AE(m-1)+(1-\alpha)FE_{selected} & \text{if} \quad AF=1 \end{cases} \tag{1}$$

식 (1)에서 m은 슈퍼 프레임 인덱스를 나타내고, $FE_{selected}$는 이전 보컬 강화 과정에서 선택된 반주 프레임의 전체 에너지를 의미한다. AF 파라미터는 이전 보컬 강화 과정에서 선택된 반주 프레임의 유, 무를 나타낸다. 식 (1)에서 AF 파라미터 값이 0이면 이전 보컬 강화 과정에서 선택된 반주 프레임이 없었음을 의미하기 때문에, 반주 신호 에너지는 이전 값을 그대로 사용한다. 반대로 AF 파라미터 값이 1이면 이전 보컬 강화 과정에서 선택된 반주 프레임이 있었음을 의미하고, 선택된 반주 프레임의 전체 에너지 $FE_{selected}$를 이용하여 $AE(m)$ 값을 결정한다. 이 때 예측된 반주 에너지 $AE(m)$은 슈퍼 프레임 단위로 갱신되기 때문에, 급격한 크기 변화를 방지하기 위해 식 (1)과 같이 이전 반주 신호의 에너지를 이용하여 스무딩 (smoothing) 처리한다. 본 논문에서는 이득 값 α의 크기를 실험을 통해 결정한 0.89로 사용했지만, 향후 입력 신호의 종류에 적응적으로 (adaptive) 값을 결정할 수 있도록 연구할 예정이다.

4) 반주 프레임 결정 동작에서는 앞에서 결정된 각 프레임의 전체 에너지 값 $FE(n)$과 슈퍼 프레임의 반주 신호 에너지 값 $AE(m)$ 을 이용하여 식 (2)와 같이 슈퍼 프레임의 각 프레임 별로 $VASNR(n)$을 계산한다.

$$VASNR(n) = \frac{FE(n)}{AE(m)} \qquad n = 0,\ 1,\ 2,\ 3,\ 4 \tag{2}$$

다섯 개의 프레임 중에서 $VASNR(n)$ 값이 가장 작은 프레임이 반주 신호의 크기가 가장 큰 것으로 판단하여 반주 복사본 신호 생성을 위한 반주 프레임으로 선택된다. 단, 다섯 프레임의 $VASNR(n)$ 값이 모두 클 경우 잘못된 반주 프레임이 선택되는 것을 방지하기 위해 미리 결정한 임계값 (threshold)을 이용한다. 즉 $VASNR(n)$이 임계값보다 크면, 현재 슈퍼 프레임에는 반주 프레임이 없는 것으로 판단하고, 다음 보컬 강화 과정에서 사용할 AF 파라미터 값을 0으로 결정한다. 반대로 $VASNR(n)$이 임계값보다 작으면, 선택된 프레임을 반주 프레임으로 최종 결정하고, AF 파라미터 값을 1로 한다.

2.3. 보컬 강화

제안한 방법의 두 번째 과정인 보컬 강화 동작 과정은 그림 4와 같다. 보컬 강화 과정에서는 선택된 반주 프레임을 이용하여 반주 복사본 신호를 생성하고, 이것을 원래 입력된 다성 음악 신호에서 제거하여 보컬 신호를 강화한다. 보컬 강화 과정의 모든 동작은 주파수 영역에서 스펙트럼 형태로 이루어지며, 각각의 세부 동작은 다음과 같다.

1) 반주 복사본 생성 동작에서는 선택된 반주 프레임의 신호 (스펙트럼)에 이득 값을 곱하여 반주 복사본 신호 $AR(m)$를 결정한다. 이 때 사용되는 이득 값 β의 크기는 0.85로 고정한다.

2) 반주 복사본 신호 비교 동작에서는 현재 슈퍼 프레임에서 결정된 반주 복사본 신호 $AR(m)$과 이전 슈퍼 프레임에서 생성되어 미리 저장해놨던 최종 반주 복사본 신호 $AR_{fianl}(m-1)$의 상관관계를 측정하여 최종 반주 복사본 신호 $AR_{final}(m)$을 결정한다. 두 신호의 상관관계가 크다고 판단되면 현재 슈퍼 프레임의 처리 과정에서 생성한 반주 복사본 신호 $AR(m)$을 최종 반주 복사본 신호 $AR_{final}(m)$으로 결정한다. 반대로 두 신호의 상관관계가 작다고 판단되면 두 신호의 평균값을 최종 반주 복사본 신호 $AR_{final}(m)$으로 결정한다. 현재 생성된 반주 복사본 신호와 과거 생성되었던 반주 복사본 신호를 비교하는 이유는 반주 복사본 신호의 급격한 변화를 최소화하여 안정적으로 보컬 강화가 이루어지도록 하기 위함이다.

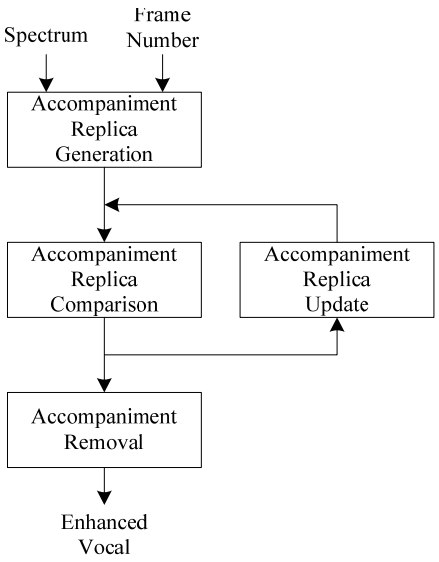

그림 4. 보컬 강화의 흐름도
Fig. 4. Flowchart of vocal enhancement.

3) 반주 복사본 신호 갱신에서는 생성된 최종 반주 복사본 신호 $AR_{final}(m)$을 저장하여 다음 슈퍼 프레임을 위한 보컬 강화 과정에서 사용할 수 있게 한다.

4) 반주 제거 동작에서는 최종 결정된 반주 복사본 신호 $AR_{final}(m)$을 입력 신호에서 제거하는데, 제거과정은 각 프레임마다 개별적으로 이루어진다. 즉, 반주 복사본 $AR_{final}(m)$은 슈퍼 프레임 단위로 정해지지만, 이를 프레임 단위로 미세 조정을 하여 프레임별로 제거할 신호를 각각 구하고 최종 제거 동작을 수행한다. 즉, 앞에서 결정한 각 프레임의 $VASNR(n)$ 값에 따라 식 (3)과 같이 프레임별로 제거할 반주 신호를 결정한다.

$$AR_{final\_modify}(m,n) = \begin{cases} 0.9\ AR_{final}(m) & VASNR(n) > th1 \\ 0.75\ AR_{final}(m) & th1 > VASNR(n) > th2 \\ 0.68\ AR_{final}(m) & th2 > VASNR(n) \end{cases} \quad (3)$$

$VASNR(n)$이 첫 번째 임계값($th1$)보다 클 경우는 보컬 신호의 크기가 매우 큰 것으로 판단하여 최종 제거할 신호를 반주 복사본 신호 $AR_{final}(m)$의 크기와 거의 동일하게 한다. 반대로 $VASNR(n)$이 두 번째 임계값($th2$)보다 작을 경우는 보컬 신호의 크기가 상대적으로 작으므로 반주 제거 과정에서 보컬 신호가 받는 영향을 줄이기 위해 $AR_{final}(m)$의 크기를 줄여 제거 신호로 사용한다. 이 과정은 반주 복사본 신호를 제거할 때 원 신호의 고유 특성이 왜곡되는 것을 줄이기 위한 것이다.

5) 최종 결정된 $AR_{final_modify}(m,n)$을 각 프레임마다 주파수 영역에서 제거하고, IDFT를 이용하여 시간 영역으로 변환하여 보컬이 강화된 신호를 구한다.

III. 성능 평가

다양한 방법을 이용하여 제안한 기술의 보컬 강화 성능을 객관적으로 검증하였다. 먼저 제안한 기술을 이용한

보컬 강화 이전과 이후 신호에 대하여 자기 상관 계수를 비교하였고, 실제 보컬 신호의 피치 주파수를 측정하여 검출 정확도 향상 정도를 측정하였다. 또한 보컬 프레임 판별 실험을 함께 진행하여 제안한 보컬 강화 기술이 보컬 피치 검출 성능 향상에 기여하는 것을 검증하였다. 실험 결과의 보편성을 위해 MIREX (music information retrieval evaluation exchange)에서 사용하는 train01 부터 train09 까지 9개의 음악 신호를 이용하여 성능을 측정하였고 [9], 모두 샘플링 주파수는 8 kHz, 모노 사운드로 고정하였다.

보컬 피치 검출은 ACF (Auto-Correlation Function)을 이용하여 기본 주파수를 정하는 방법이 널리 사용된다. 따라서 보컬 강화 이전과 이후의 자기 상관 계수의 변화를 분석하여 보컬 성분이 강화된 것을 확인할 수 있다. 그림 5 (a)는 오리지널 train01 데이터의 360번째 프레임에 대한 자기 상관 계수를 나타낸 것으로, 56번째 인덱스에서 가장 큰 자기 상관 계수 값을 가진다. 전처리를 해준 train01 데이터의 360번째 프레임에 대한 자기 상관 계수를 나타낸 그림 5 (b)에서는 27번째 인덱스에서 가장 큰 자기 상관 계수 값을 갖는 것을 알 수 있다. 이렇게 측정된 두 개의 자기 상관 계수 결과를 이용하여 360번째 프레임의 피치 주파수 값을 계산하면 오리지널 신호의 경우는 약 142.85 Hz,

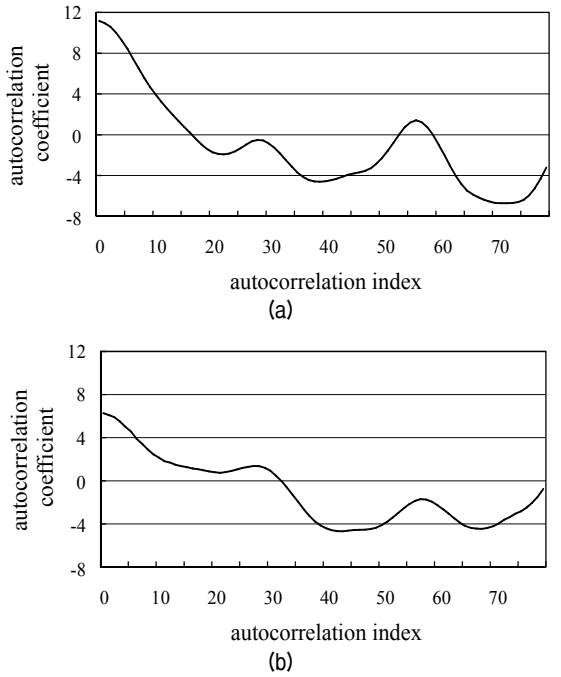

그림 5. train01 데이터 360번째 프레임의 자기 상관 계수
(a) 오리지널 신호 (b) 전처리를 한 신호

Fig. 5. Autocorrelation coefficient of train01 data at 360th frame.
(a) Original signal (b) Preprocessed signal

전처리된 신호의 경우는 약 296.29 Hz의 피치 주파수 값이 결정된다. train01 데이터의 360번째 프레임에 대한 레퍼런스 피치 값은 294.03 Hz이기 때문에 제안한 보컬 강화 기술을 이용하여 전처리를 했을 경우에만 정확한 보컬 피치를 검출되었음을 확인할 수 있다. 실제 train01 신호의 360번째 프레임에는 여성의 음성이 피아노 반주와 함께 섞여있기 때문에 제안한 강화 기술로 전처리를 하지 않을 경우, 보컬 신호보다 좀 더 하모닉 특성이 뚜렷한 피아노 소리의 피치 값이 검출된다.

제안한 보컬 강화 기술을 사용하여 실제 보컬 피치 검출의 정확도가 향상되는지를 검사하였다. 보컬 피치를 검출하는 실험에서는 일반적으로 많이 사용되는 자기 상관 계수를 이용한 피치 주파수 측정 방식과 Li와 Wang 기술 [10]을 사용하는 피치 주파수 측정 방법을 각각 이용하였다. Li와 Wang 기술은 다중 피치 검출 기술을 기반으로 보컬 피치를 검출하며, 자기 상관 함수와는 무관한 방법이다. 보컬 프레임과 반주 프레임을 구별하는 실험에서는 다성 음악 신호의 단기간 에너지 (short-term energy) 특성을 이용한 측정 방식을 사용하였다 [11]. 성능 검증에서 사용한 항목은 다음과 같다.

1) Vocal Pitch Accuracy (only vocal frame) : MIREX에서 제공한 레퍼런스 (reference) 피치 주파수 값과 보컬 피치 검출 방법을 이용하여 검출한 피치 주파수 값을 비교하여, 일치한 개수를 백분율로 나타낸 성능이다. 이 때 주파수 일치 여부는 주파수 값의 차이가 ± 1/4 톤 (tone) 이내일 경우로 규정한다. 성능 측정은 전체 프레임 중에서 레퍼런스에서 결정한 보컬 프레임에서만 진행한다.

2) Vocal Pitch Accuracy (overall) : 위의 Vocal Pitch Accuracy와 동일한 방법으로 피치 정확도를 측정하고, 단지 성능 측정을 모든 프레임 (보컬 프레임+반주 프레임)에서 진행한다.

3) Vocal Frame Detection : 데이터의 실제 보컬 프레임 중에서 보컬 프레임 판별법을 통하여 보컬 프레임으로 판별된 프레임의 개수를 백분율로 표현한다. 이 때 보컬 프레임의 여부는 MIREX에서 제공한 레퍼런스 문서를 참조한다.

4) Vocal False Alarm : MIREX에서 제공한 레퍼런스 문서를 참조하여, 실제 보컬 프레임이 아니지만 보컬 프레임 판별법을 통해 보컬 프레임으로 잘못 판단한 프레임의 개수를 백분율로 표현한다.

표 1과 표 2는 MIREX에서 사용하는 9개의 다성 음악 신호에 대해서 위의 네 가지 항목을 측정한 실험의 결과를 보여준다. 먼저 표 1의 Vocal Pitch Accuracy (overall)

표 1. 보컬 피치 검출 성능 비교
Table 1. Comparison of vocal pitch accuracy.

	Vocal Pitch Accuracy (only vocal frame)				Vocal Pitch Accuracy (overall)			
	ACF		Li & Wang		ACF		Li & Wang	
	Original	Pre-processed	Original	Pre-processed	Original	Pre-processed	Original	Pre-processed
train01	80.9 %	90.4 %	83.1 %	91.7 %	73.6 %	84.3 %	78.8 %	88.4 %
train02	65.6 %	71.1 %	74.6 %	80.5 %	61.9 %	68.5 %	70.1 %	77.9 %
train03	59.8 %	63.7 %	70.2 %	78.1 %	63.1 %	70.2 %	68.7 %	75.2 %
train04	58.1 %	67.4 %	72.1 %	77.9 %	56.7 %	63.6 %	70.5 %	76.8 %
train05	55.2 %	62.5 %	68.5 %	72.4 %	52.3 %	60.6 %	65.8 %	70.1 %
train06	60.5 %	65.3 %	74.8 %	78.3 %	58.8 %	64.9 %	69.4 %	75.6 %
train07	65.4 %	71.9 %	79.1 %	82.6 %	65.0 %	69.1 %	75.2 %	79.8 %
train08	55.5 %	62.4 %	70.6 %	74.2 %	54.7 %	61.2 %	68.9 %	71.5 %
train09	61.7 %	69.1 %	75.4 %	80.9 %	60.2 %	67.7 %	71.3 %	78.3 %
Ave.	62.5 %	69.3 %	74.3 %	79.6 %	60.7 %	67.8 %	71.0 %	77.1 %

표 2. 보컬 프레임 판정 정확도 비교
Table 2. Comparison of vocal frame detection accuracy.

	Vocal Frame Detection		Vocal False Alarm	
	Original	Pre-processed	Original	Pre-processed
train01	84.5 %	87.2 %	40.5 %	27.6 %
train02	83.9 %	79.1 %	45.3 %	40.0 %
train03	75.7 %	73.8 %	31.9 %	21.3 %
train04	60.5 %	61.7 %	36.9 %	28.0 %
train05	50.4 %	63.6 %	39.2 %	36.2 %
train06	80.2 %	79.0 %	47.8 %	35.2 %
train07	77.6 %	84.7 %	33.8 %	39.5 %
train08	66.3 %	64.8 %	43.2 %	33.7 %
train09	74.3 %	73.9 %	44.4 %	30.1 %
Ave.	72.6 %	74.2 %	40.3 %	32.4 %

항목을 보면 오리지널 신호에 대한 보컬 피치 정확도보다 제안한 보컬 강화 기술을 이용해 전처리된 신호에서 검출한 보컬 피치의 정확도가 ACF 방법과 Li 와 Wang 방법에서 각각 평균 7.1 % 포인트와 6.1 % 포인트 향상되었음을 알 수 있다. 특히 Vocal Pitch Accuracy (only vocal frame) 항목의 결과는 보컬 프레임에서 보컬 프레임의 피치 값을 정확히 검출했음을 보여주는 결과이다. 다만 ACF 방법에서 train03 데이터의 경우 Vocal Pitch Accuracy (only vocal frame) 값이 Vocal Pitch Accuracy (overall) 값보다 작게 나타난다. 이는 다른 데이터들에 비해 train03 데이터의 보컬 프레임 부분이 뭉쳐있지 않고 산재해 있기 때문에 발생한 현상으로 해석된다. 즉 반주 프레임에서 보컬 프레임으로 넘어가는 부분과 반대로 보컬 프레임에 반주 프레임으로 넘어가는 부분에서 일반적으로 보컬 피치 검출의 오류가 많이 발생하는데, 이러한 부분들이 다른 데이터에 비해 train03에 많기 때문이다. 표 1을 통하여 오리지널 신호보다 제안한 보컬 강화 기술을 이용해 전처리된 신호에서 보컬 피치 검출의 정확도가 더 높은 것을 검증하였고, 특히 기존의 보컬 피치 검출 방법을 그대로 사용하고 제안한 보컬 강화 기술만을 통하여 보컬 피치 검출의 정확도가 크게 향상되는 것을 확인할 수 있다.

표 2는 두 개의 다성 음악 신호를 보컬 프레임 (보컬 피치가 있는 프레임)과 반주 프레임 (보컬 피치가 없는 프레임)으로 구별하는 실험의 결과를 보여준다. 보컬 프레임 판정의 정확도는 실제 보컬 피치 검출의 정확도 향상과 밀접한 관계를 갖고 있기 때문에 매우 중요한 항목이다. 표 2를 통하여 제안한 보컬 강화 기술에 의하여 Vocal Frame Detection 평균값이 약간 증가하고, 특히 Vocal False Alarm 평균값이 많이 작아졌음을 알 수 있다. 이러한 결과는 보컬 프레임 판별 시 발생하는 오차의 원인이 되는 반주 신호가 제안한 보컬 강화 기술에 의해 효과적으로 제거되었음을 의미한다. 즉 제안한 보컬 강화 기술에 의해 음악 신호에서 반주 신호가 정확히 제거되기 때문에 보컬 프레임을 좀 더 정확히 찾게 되고, 결과적으로 보컬 신호가 있는 것으로 잘못 판단된 프레임의 숫자가 확연히 줄어들 뿐만 아니라 보컬 피치 검출의 정확도 향상에도 도움을 주게 된다. 따라서 제안한 보컬 강화 기술은 보컬 피치를 정확히 찾는데 기여할 뿐만 아니라 보컬 프레임과 반주 프레임을 판정하는 성능도 향상시키는 것을 확인 할 수 있다.

이상의 성능 평가 결과를 통하여 제안한 보컬 강화 기술에 의하여 입력 신호의 보컬 신호가 실제로 강화되었고, 그 결과 보컬 피치 검출의 정확도와 보컬 프레임 판정

의 정확도가 향상된 것을 검증하였다. 특히 제안한 기술을 활용하면 기존의 다양한 방법으로 개발된 보컬 피치 검출 방법과 보컬 프레임 판정 방법들을 그대로 사용하면서 각각의 성능을 향상시킬 수 있으므로, 제안 기술은 보컬 신호를 검출하는 다양한 응용 분야에 적용될 수 있을 것이다.

IV. 결 론

본 논문에서는 다성 음악 신호의 보컬 피치 검출의 성능 향상을 위한 보컬 강화 기술을 제안하였다. 제안한 기술은 잡음 억제 알고리즘을 기반으로 입력된 다성 음악 신호에서 반주 신호를 예측하고, 가공하여 원래 다성 음악 신호에서 제거한다. 이 때 제거되는 반주 신호에 의해 잔여 반주 신호의 하모닉 특성은 손상되고, 좀 더 잡음 특성을 갖게 된다. 이러한 결과는 보컬 피치 검출의 정확도 향상에 기여한다. 제안한 기술은 슈퍼 프레임 단위로 반주 성분이 강한 프레임을 선정하고, 이를 기준으로 반주 복사본 신호를 결정하고, 신호의 특성에 따라 미세 조정을 거쳐 프레임 단위로 입력 신호에서 반주를 제거한다.

보컬 피치 검출 실험과 보컬 프레임 판별 실험을 통해 제안한 보컬 강화 기술에 의하여 보컬 신호 검출과 보컬 프레임 판정의 정확도가 향상되는지를 분석하였고, 제안한 기술에 의하여 보컬 피치의 정확도가 평균 7.1 % 포인트 향상되었음을 확인하였다. 또한 전혀 다른 두 개의 보컬 피치 검출 기술을 이용하여 보컬 피치 정확도를 평가한 결과, 두 경우 모두 제안한 기술에 의해 보컬 피치 검출의 정확도가 향상되었음을 확인할 수 있었다. 이를 통하여 제안한 보컬 강화 기술에 의해 반주 신호가 효과적으로 제거 되고, 기존의 다양한 보컬 피치 검출 방법을 그대로 사용하여도 보컬 피치 검출 성능 향상에 기여할 수 있다는 것을 검증하였다. 단, 제안한 보컬 강화 기술은 다성 음악 신호에 포함된 반주 신호를 제거하여 보컬 신호를 강화하는 기술이므로 보컬 신호가 없는 다성 음악 신호에 적용하면 오히려 반주 신호를 훼손하여 정확한 반주 피치를 찾는 것이 어려워진다. 보컬 신호가 없는 다성 음악 신호의 경우에도 메인 멜로디의 피치를 정확히 찾을 수 있는 연구가 추가 진행될 예정이다.

감사의 글

이 논문은 2011년도 광운대학교 연구년에 의하여 연구되었습니다.

참 고 문 헌

1. Yipeng Li and DeLiang Wang, "Detecting pitch of singing voice in polyphonic audio," *IEEE Conf. Acoustics, Speech, and Signal Processing*, vol. 3, pp. 17–20, 2005.
2. Jean–Louis Durrieu, Gael Richard and Bertrand David, "Singer melody extraction in polyphonic signals using source separation methods," *IEEE Conf. Acoustics, Speech, and Signal Processing*, vol. 43, no. 4, pp. 169–172, 2008.
3. Masataka Goto, Takeshi Saitou, Tomoyasu Nakano and Hiromasa Fujihara, "Singing Information Processing based on singing voice modeling," *IEEE Conf. Acoustics, Speech, and Signal Processing*, pp. 5506–5509, 2010.
4. Vishweshwara Rao and Preeti Rao, "Vocal melody extraction in the presence of pitched accompaniment in polyphonic music," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, pp. 2145–2154, 2010.
5. Anssi Klapuri, "Multipitch Analysis of Polyphonic Music and Speech Signals Using an Auditory Model." *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, pp. 255–266, 2008.
6. N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka and S. Sagayama "Separation of a monaural audio signals into harmonic/percussive components by complementary diffusion on spectrogram," *Processings of EUSIPCO*, 2008.
7. TIA/EIA/IS–127, Enhanced Variable Rate Codec, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems, Jan. 1997.
8. S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech, Signal Processing,* vol. 27, pp. 113–120, 1979.
9. http://labrosa.ee.columbia.edu/projects/melody
10. Yipeng Li and DeLiang Wang, "Separation of singing voice from music accompaniment for monaural recording," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, pp. 1475–1487, 2007.
11. Sen Zhang, "An energy–based adaptive voice detection approach," *Proc. 8th International Conf. Signal Processing*, vol. 1, pp. 1109–1113, 2006.

저자 약력

●이 세 원 (Sewon Lee)
한국음향학회지 제30권 제3호 참조

●송 재 종 (Chai–Jong Song)
1999년: 원광대학교 전자공학과 (공학사)
2001년: 광운대학교 전자공학과 (공학석사)
2001년 ~ 현재: 전자부품연구원 연구원
※ 주관심 분야 : 음성/오디오 신호처리, 음악신호 분석

●이 석 필 (Seok–Pil Lee)
1990년: 연세대학교 전기공학과 (공학사)
1992년: 연세대학교 전기공학과 (공학석사)
1997년: 연세대학교 전기공학과 (공학박사)
1997년 ~ 2002년: 대우전자 영상 연구소 선임연구원
2002년 ~ 현재: 전자부품연구원 디지털미디어연구센터 센터장
※ 주관심 분야 : 디지털방송통신융합시스템, 멀티미디어 신호처리

●박 호 종 (Hochong Park)
한국음향학회지 제30권 제3호 참조