일반논문-08-13-3-09

음성 특성을 고려한 가라오케 시스템

김 유 승[a], 김 인 철[a]‡

A Karaoke system based on the vocal characteristics

Yu Seung Kim[a], and Rin Chul Kim[a]‡

요 약

본 논문에서는 음성 특성에 기반을 둔 보컬 영역 검색 알고리듬을 적용하는 가라오케 시스템을 제시한다. 제안한 시스템에서 입력 음악은 보컬 영역 검색 알고리듬을 통해 보컬 부분과 반주 부분으로 분류된다. 그런 다음, 보컬 영역에 대해서만 보컬 제거 기법을 적용한다. 보컬 영역 검색에서는 TICFT (twice iterated composite Fourier transform) 영역에서 보컬의 특성을 고려하여 분류를 수행한다. 보컬 제거를 위해서 대역 통과 필터링 된 보컬 영역으로부터 보컬 성분을 추출하고, 이를 원래의 음악에서 감산함으로써 보컬 성분이 제거된 음악을 얻는다. 본 논문에서 제시한 기법은 4곡의 노래에 적용하고, 그 성능을 평가한다.

Abstract

This paper presents a karaoke system employing a vocal region detection algorithm based on the vocal characteristics. In the proposed system, an input song is classified into vocal and instrumental regions using the vocal region detection algorithm. Then, a vocal removal method is applied only to the vocal region. To detect vocal region, a classification algorithm is designed based on the vocal characteristics in the TICFT (twice iterated composite Fourier transform) domain. For vocal removal, vocal components are extracted from a band pass filtered vocal region and they are subtracted from the original song, yielding a vocal removed song. The performance of the proposed method is measured on four different songs.

Keywords : vocal removal, vocal region detection, twice iterated composite Fourier transform

Ⅰ. 서 론

노래 반주기 등의 가라오케 (karaoke) 시스템[1]은 반주 (instrumental)와 보컬 (vocal)을 별도의 트랙에 저장한 다음, 사용자의 요청에 따라 반주만을 재생할 수 있는 시스템을 말한다. 만약 일반 CD (compact disk) 등에 수록된 노래들을 그대로 이용할 수 있다면 PC 등에서 가라오케 기능을 손쉽게 이용할 수 있다. 가라오케 전용 노래 대신에 일반 노래를 이용하는 가라오케 시스템에서는 노래에서 보컬(vocal) 성분을 분리하여 반주 성분만을 추출하는 보컬 제거 (vocal removal) 기법[2]에 기반을 둔다. 현재 보컬 제거 방법은 오디오 편집 프로그램인 Adobe사의 Audition[3]과 MS Windows XP Media Center Edition의 Vocal Remover[4] 등에서 제공되고 있다. 그리고 Cyber Link사의 Power DVD, Gretech사의 Gom player, Kmplayer, AnalogX사에서 나온 WinAmp등 범용 멀티미디어 재생기들도 보컬 제거 기법을 적용하고 있다.

기본적인 보컬 제거 기법은 노래가 스튜디오에서 녹음될

a) 서울시립대학교 전자전기컴퓨터공학부
Dept. of Electrical & Computer Eng., University of Seoul
‡ 교신저자 김인철 (rin@uos.ac.kr)
※ 이 논문은 2006년도 서울시립대 학술연구조성비에 의하여 연구되었음.

때 반주는 스테레오 (stereo)로 녹음되고 보컬은 모노 (mono)처럼 좌우 채널이 동일하게 녹음 된다는 특징을 이용한다. 이때 좌우 채널을 평균함으로써 보컬 성분을 추출하고, 이를 양 채널에서 감산하여 반주 성분만 추출할 수 있다. 보다 효과적인 보컬 성분 제거를 위해서 좌우 채널의 모든 대역을 감산하는 것이 아니라 보컬 성분의 에너지가 집중된 대역에 대해서만 감산하는 방법도 소개되어 있다[2]. 그러나 이러한 방법은 가수들의 특성이나 곡의 특성을 고려하지 않고 고정된 대역의 대역통과 필터(250Hz~9000Hz)로 필터링 된 데이터들을 감산함으로써 보컬을 분리하였다[1]. 하지만 보컬 성분이 없는 구간에서도 불필요한 감산을 수행하므로 반주 부분에 왜곡이 발생된다. 이러한 왜곡을 최소한으로 줄이기 위해서는 보컬 영역 검색이 필요하다.

보컬 영역 검색[5,6]은 음성인식 및 음악 분류를 위해 연구되던 분야로서 입력되는 소리에서 보컬 영역만을 분리하는 기법이다. 이 기법을 보컬 제거에 적용한다면, 노래에서 반주로만 구성된 구간과 보컬 성분이 포함되어 있는 구간으로 분리할 수 있을 것이다. 따라서 보컬 영역에 대해서는 보컬 제거 기법을 적용하되, 반주 부분은 원음을 그대로 출력하여 보컬 제거 과정에서 발생되는 반주 왜곡 현상을 최소한으로 줄일 수 있을 것이다.

이에 따라, 본 논문에서는 보컬 영역 검색을 통해 보컬 부분과 반주 부분을 분류한 다음, 보컬 영역에 대해서만 보컬 제거 기법을 적용하는 가라오케 시스템을 제안한다. 제안한 기법에서는 TICFT (twice iterated composite Fourier transform)에 기반을 둔 보컬 영역 검색 기법[6]을 채용하여 보컬 영역을 검색한다. 본 논문에서는 TICFT 영역에서 검색을 수행하되, 음성 특성을 고려한 분류를 수행함으로써 보다 효율적인 보컬 영역 검색이 이루어지도록 한다. 그런 다음, 검색된 보컬 영역에 대해서 대역통과 필터를 이용한 보컬 제거[2]를 수행한다. 본 논문에서 제시한 기법들을 4곡의 노래에 적용하여 그 성능을 평가한다.

본 논문은 서론에 이어 제2절에서는 TICFT 기반 보컬 영역 검색 기법에 대하여 설명하고, 제3절에서는 제안한 보컬 제거 기법에 대하여 설명한다. 제4절에서는 실험 결과를 제시하고, 제5절에서는 본 논문의 결론을 맺겠다.

II. TICFT 기반 보컬 영역 검색

본 절에서는 TICFT에 기반한 보컬 영역 검색 기법[6]에 대하여 설명한다. 그림 1에 제시한 바와 같이 보컬 영역 검색 기법은 리듬 추출 (rhythm extraction), 4분 음표 길이로의 프레임 구성 (quarter note level framing), TICFT 및 부대역 에너지 계산, 그리고 문턱값 비교 (thresholding)등의 과정을 거쳐 보컬 영역과 반주 영역을 분리한다. 이제 그림 1의 TICFT 기반 보컬 영역 검색 기법을 좀 더 구체적으로 살펴보겠다.

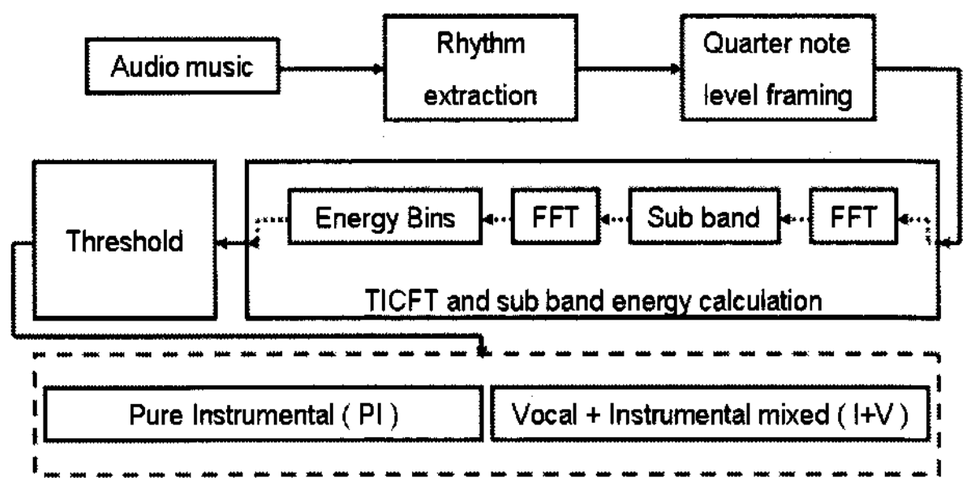


그림 1. 보컬 영역 검색 기법의 블록선도[6]
Fig. 1. A block diagram of the vocal region detection method[6]

먼저 입력된 오디오 신호는 리듬 추출 블록에서 입력 신호의 리듬을 분석하여 리듬의 기본 빠르기인 4분 음표 길이를 구하게 된다. 여기서 측정된 4분 음표 길이를 기본 단위로 입력된 오디오 신호를 시간영역에서 각각의 세부 프레임으로 나누고, 이 이후에는 프레임 단위로 처리한다. 4분 음표 길이로 프레임을 나누면, 프레임 내에서는 하모닉 구조가 quasi-stationary 한 것으로 볼 수 있어, 분석이 매우 용이해진다.

보컬 영역 검색 시스템의 핵심인 TICFT 블록에서는 나누어진 세부 프레임 단위로 FFT를 2회 반복 실행한 다음, 부대역 (subband) 에너지를 구한다. 여기서, 사람 귀의 주파수 해상도는 3Hz인 것으로 알려져 있으므로[7], FFT의 해상도는 3Hz 이하로 설정한다. 본 논문에서는 표본화 주파수가 44.1KHz인 오디오에 대해서 FFT 프레임 길이를 32,768 샘플로 하여 주파수 해상도가 1.35Hz가 되도록 한다.

일반적으로, 반주 성분은 하모닉 (harmonic) 구조가 전체

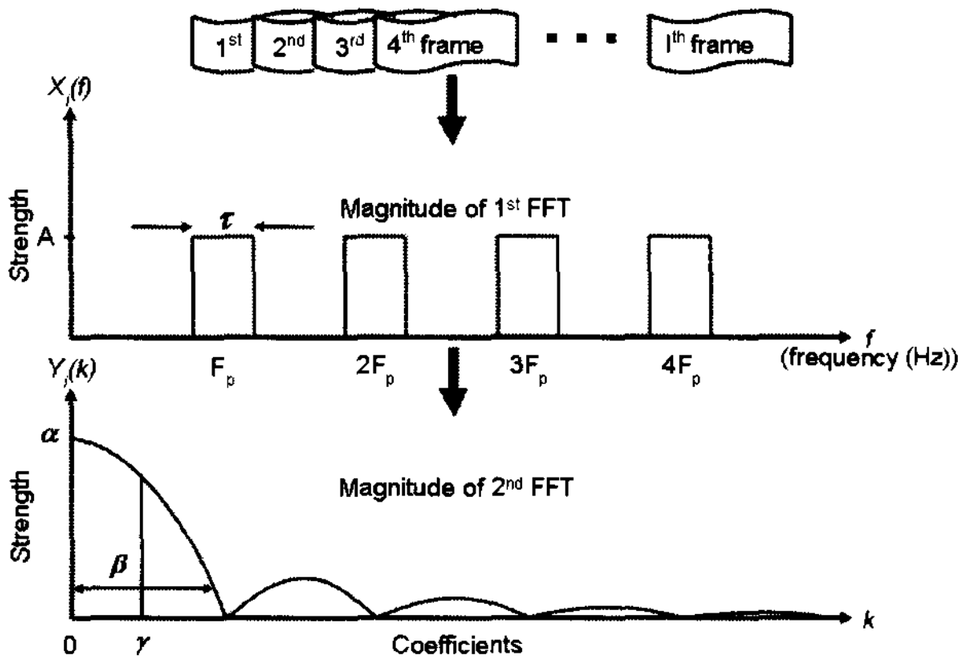


그림 2. TICFT[b]
Fig. 2. TICFT[6]

주파수 공간에 고루 분포하는 것에 반해 보컬 성분은 주파수 영역의 좁은 대역에서 강한 피치 (pitch) 하모닉 구조를 가진다고 알려져 있다[8]. 이러한 점을 참조하여 TICFT에서는 먼저 첫 번째 FFT를 실행한 다음, FFT 영역에서 통과 대역이 512~2048Hz인 필터로 대역통과 필터링한다. 하모닉 구조를 가지는 신호는 FFT 영역에서 크기 스펙트럼 (magnitude spectrum)이 그림 2에 보인바와 같이 주기 함수로 표현된다. 즉, 첫 번째 FFT 후 크기 스펙트럼(magnitude spectrum)은 주기(period)가 F_p, 펄스 폭 (pulse width)이 τ, 진폭(amplitude)이 A인 주파수 펄스열 (frequency pulse train) $X(f)$로 표현된다. 이 $X(f)$에 대역통과 필터와 두 번째 FFT를 거친 다음 절대값을 취하면 다음과 같이 sinc 함수와 같은 형태의 신호가 산출된다.

$$Y_i(k) = \left| \frac{A\tau}{F_p} \frac{\sin \pi k F_o \tau}{\pi k F_o \tau} \right|. \quad (1)$$

여기서 $Y_i(k)$는 i 번째 프레임에서 k번째 TICFT 계수의 크기를 말하고, F_o는 TICFT 계수 사이의 간격을 나타낸다. 식 (1)과 그림 2의 아래 그림을 참조하면, $Y_i(k)$의 주 돌출부 (main lobe) 높이 α는 τ/F_p에 비례하고, 폭 β는 τ에 반비례한다.

다음으로, 그림 1의 문턱값 비교 블록에서는 식 (2)와 같이 각 프레임별로 γ번째 계수까지 $Y_i(k)$의 합 $B_l(i)$를 얻는다.

$$B_l(i) = \sum_{k=1}^{\gamma} Y_i(k). \quad (2)$$

그런 다음, $B_l(i)$에서 $B_l(i)$의 평균을 제거함으로써 $B_l(i)$를 정규화 시킨 $B_l^{MR}(i)$를 다음과 같이 얻는다.

$$B_l^{MR}(i) = B_l(i) - \frac{1}{n}\sum_{i=1}^{n} B_l(i). \quad (3)$$

문턱값 비교 블록에서는 식 (3)의 $B_l^{MR}(i)$ 값으로 보컬 프레임과 반주 프레임을 분류한다. 만약 입력된 프레임에서 $B_l^{MR}(i) > 0$이라면 그 프레임은 보컬성분이 포함된 프레임으로 분류한다. 반면에 $B_l^{MR}(i) \le 0$이라면 그 프레임을 반주 성분으로만 구성된 프레임으로 분류한다.

III. 제안한 보컬 제거 기법

본 절에서는 제안하는 보컬 제거 기법에 대하여 설명한다. 제안하는 기법에서는 먼저, II절에서 설명한 기본적인 TICFT 기반 보컬 영역 검색 기법을 수정하여 그 성능을 향상시킬 수 있는 보컬 영역 검색 기법을 제시한다. 그런 다음, 검색된 보컬 영역에 대해서만 보컬 제거를 수행하는 새로운 보컬 제거 기법을 제안한다.

1. 수정된 보컬 영역 검색

여기서는 II절에서 설명한 기법의 성능을 개선할 수 있도록 수정된 TICFT 기반 보컬 영역 검색 기법을 제시한다. 수정된 기법에서는 II절에서 설명한 기법과 마찬가지로 TICFT를 이용하되 부대역 에너지를 구하는 방법을 수정하고 후처리 과정을 새로이 도입하여, 보다 정확하게 보컬 영역을 검색하도록 한다.

먼저, TICFT 계수의 크기의 합으로 부대역 에너지를 얻는 식 (2)를 다음과 같이 R1부터 R2까지의 TICFT 계수의

크기를 합하는 것으로 합하는 범위를 제한하였다.

$$B_l(i) = \sum_{k=R1}^{R2} Y_i(k). \quad (4)$$

즉, TICFT에서는 첫 번째 FFT 후 크기 스펙트럼에 대해 두 번째 FFT를 수행하므로, $Y_i(k)$의 저주파 성분은 모든 프레임에 대해 매우 큰 값을 가지게 된다. 따라서 R1은 1보다 큰 값으로 설정하여야 하며, 본 논문에서는 5로 하였다. 또한, R2는 식 (2)에서 γ와 같은 값으로 보컬의 경우 낮은 위치의 계수들이 매우 큰 값을 가지는 것으로 나타났다. 본 논문에서 R2는 100으로 설정하였다.

두 번째로, 식 (4)의 $B_l(i)$에 길이가 5인 메디안 필터를 적용하여 임펄스성 잡음을 제거하였다. 그런 다음, 보컬 영역 및 반주 영역에서의 $B_l(i)$의 국부적인 평균값을 구한다. 즉, 만약 i 번째 프레임이 보컬 영역이라면 보컬 평균을 다음과 같이 갱신하고,

$$M_{vocal} = \alpha \times M_{vocal} + (1-\alpha) \times B_l(i), \quad (5)$$

반주 영역이라면 반주 평균을 다음과 같이 갱신한다.

$$M_{inst} = \alpha \times M_{inst} + (1-\alpha) \times B_l(i). \quad (6)$$

그런 다음, 다음 프레임의 $B_l(i+1)$이 보컬 평균과 반주 평균의 평균값보다 크면 그 프레임을 보컬 영역으로, 그렇지 않으면 반주 영역으로 판별한다. 이때, α는 0.95로 설정하였다. 그리고 보컬 평균의 초기값은 전체 곡에 대한 $B_l(i)$의 평균값으로 설정하고, 반주 평균의 초기값은 최초 20 프레임에 대한 $B_l(i)$의 평균으로 설정한다.

식 (5)와 (6)의 평균으로 문턱값을 설정하는 것은 영역을 나누지 않고 얻는 단순한 $B_l(i)$의 평균과는 차이가 있다. 먼저, 이 방법을 통해 오디오의 국부적인 특성을 고려할 수 있게 된다. 즉, $B_l(i)$ 값의 크기에 따라 문턱값이 조정되어 큰 크기의 보컬이 있는 영역과 작은 크기의 보컬이 있는 영역에 대해 적응적으로 영역을 결정할 수 있다. 앞에서 $B_l(i)$에 대해 메디안 필터를 적용한 것은 각 영역에 대한 국부적인 평균이 지나치게 변하는 것을 방지하여 안정적으로 문턱값이 설정될 수 있게 도와준다. 그리고 보컬 영역과 반주 영역에 대한 개별적 평균을 얻음으로써 각 영역의 국부적인 특성을 적절히 반영할 수 있게 된다.

세 번째로, 고립된 보컬 프레임 혹은 반주 프레임은 잘못 판별되었을 가능성이 높으므로, 메디안 필터링을 통해 고립된 영역들을 제거한다. 본 논문에서는 7차의 메디안 필터를 적용하여, 3개 이하의 고립된 프레임들을 제거하였다.

2. 보컬 제거

그림 3에서는 제안하는 보컬 제거 기법의 블록선도를 제시하였다. 제안하는 보컬 제거 기법에서는 보컬 영역 검색 기법을 적용하고, 보컬 영역으로 검색된 프레임에 대해서 보컬 제거 기법을 적용한다. 그림 3을 참조하면서 제안한 시스템을 보다 구체적으로 살펴보면 다음과 같다.

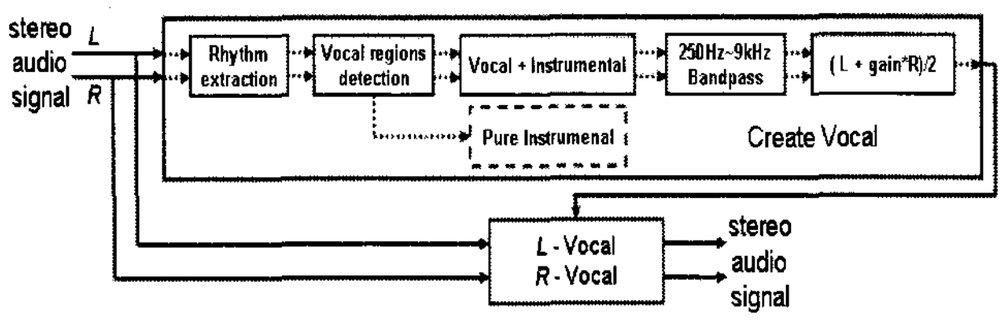


그림 3. 제안하는 보컬 제거 기법의 블록선도
Fig. 3. Block diagram of the proposed vocal removal method

먼저, 리듬 추출 블록에서는 입력된 오디오 신호를 50% 오버랩(overlap) 시켜 32,768 샘플 단위의 프레임으로 나눈다. 즉, 본 논문에서는 약 0.74초라는 고정된 크기로 4분음표 길이를 설정한다. 나누어진 각 프레임은 보컬 영역 검색 블록에서 보컬 프레임과 반주 프레임으로 분류된다. 이때 보컬 프레임으로 판별된 프레임에 대해서 250Hz~9kHz의 대역통과 필터로 필터링 한다. 그런 다음, 식 (7)과 같이 필터링 된 좌우 채널 신호의 가중합을 취함으로써 보컬 신호를 산출한다.

$$Vocal = \frac{1}{2} \times (L + gain \times R). \quad (7)$$

여기서 L과 R은 각각 필터링 된 좌측 및 우측 채널 신호이다. $gain$은 식 (8)과 같이 양 채널 사이의 에너지 비를 구하고,

$$\sum(L)^2 = k \times \sum(R)^2, \tag{8}$$

이를 식 (9)와 같이 크기 비로 나타낸 것이다.

$$gain = \sqrt{k}. \tag{9}$$

마지막으로, 입력된 오디오 신호의 좌우 채널로부터 식 (7)과 같이 생성된 보컬을 감산하면 보컬이 제거되고 반주 성분만이 남은 오디오 신호를 얻을 수 있다.

한편, 제안한 보컬 제거 기법에서는 보컬 영역으로 검색된 프레임에 대해서만 보컬 성분을 제거한다. 이때 보컬 프레임의 가장자리에서는 보컬을 제거하는 처리의 절환으로 인해 출력이 급격히 변하게 된다. 이러한 단점을 해결하기 위해, 보컬 프레임의 전후에 위치한 반주 프레임은 별도로 처리하여, 보컬 제거 절환이 서서히 이루어지게 한다. 즉, 보컬 프레임의 앞에 위치한 반주 프레임에는 식 (7)에 의해 추출된 보컬 성분에 32,768 길이의 Hann 윈도우 전반부를 취한 것을 좌우 채널에서 감산한다. 그리고 보컬 프레임의 뒤에 위치한 반주프레임에는 보컬 성분에 Hann 윈도우 후반부를 취한 것을 좌우 채널로부터 감산한다. 그 결과 보다 자연스러운 출력을 얻을 수 있도록 한다.

Ⅳ. 실 험

본 절에서는 제안한 보컬 제거 기법의 성능을 평가하기 위해 4곡의 노래를 이용하여 테스트 하였다. 실험에 사용한 노래는 유미의 “Knocking on heaven's door”와 이승철의 “희야”, 보아의 "Moon & sunrise", 그리고 한국 여성 보컬(미상)의 “If I ain't got you" 이다. 이 중 앞 3곡은 CD에 수록된 것을 44.1KHz로 표본화 하였고, 샘플 당 16 비트로 표현하여 wave 파일 형식으로 컴퓨터에 저장하여 사용하였다. 그리고 마지막 곡은 MP3로 압축된 곡이다. 유미와 보아, 미상은 여성 보컬이고, 이승철은 남성 보컬이다. 4곡 모두 흠 (humming), 화음 등 백보컬 (backing vocals)이 포함되어 있고, 특히 보아의 곡은 백보컬이 많은 부분을 차지한다. 또한, 유미의 곡은 보컬의 세기가 큰 변화가 없지만, 다른 3곡은 곡 전반에 걸쳐 보컬의 세기가 상당히 변하는 특징을 보인다. 본 절에서는 먼저 보컬 검색 기법에 대한 실험 결과를 서술하고, 보컬 제거 기법에 대한 설명을 하겠다.

1. 보컬 검색 실험

여기서는 앞에서 소개한 4곡의 실험 노래를 이용하여 보컬 검색 기법의 성능을 평가하겠다. 먼저, 이 wave 파일들을 16384 샘플로 구성되는 프레임 단위로 보컬 영역과 반주 영역으로 판별하였다. 보컬 및 반주 영역 판별은 직접 귀로 반복 청취하여 실험 노래 각각에 대하여 그림 4와 같은 보컬 영역 테이블을 만들고, 이를 실험 결과와 비교하였다. 보컬 영역 테이블을 작성한 결과, 유미의 곡은 총 568프레임으로 이 중 보컬은 210프레임이며 반주는 358프레임이다. 이승철의 곡은 총 396프레임으로 보컬은 227프레임이며 반주는 169프레임이다. 보아의 곡은 총 841프레임으로 보컬은 681프레임이고 반주는 160프레임이다. 마지막으로 미상의 곡은 총 606프레임으로 보컬은 469프레임이고 반주는 137프레임이다.

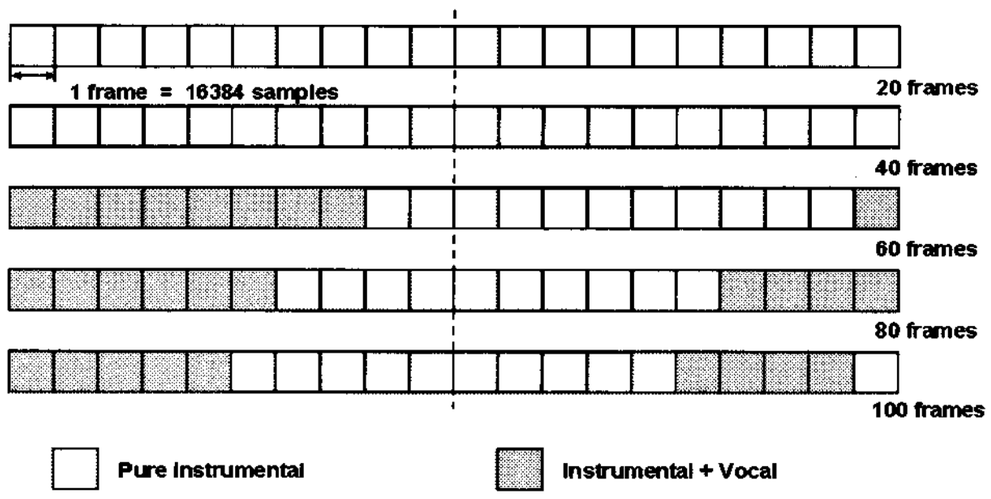


그림 4. 유미 노래에 대한 보컬 및 반주 영역 분류 테이블
Fig. 4. A classification table of vocal and instrumental regions for the song by Yumi

이 보컬 영역 테이블을 기준으로 하여 보컬 영역 검색 기법의 정확도를 측정한다. 정확도는 hit, miss, false alarm

을 정의하여 측정하였다. 그림 5에서는 3가지 경우의 예를 제시하였다. Hit는 보컬 프레임을 정확하게 보컬 프레임으로 판별한 것으로, hit ratio는 전체 보컬 프레임 개수에 대한 hit 프레임 개수의 비이다. Miss는 보컬 프레임을 반주 프레임으로 잘못 판별한 것을 말하며, miss ratio는 전체 보컬 프레임 개수에 대한 miss 프레임 개수의 비이다. False alarm은 반주 프레임을 보컬 프레임으로 잘못 판별한 것을 말하며, false alarm ratio는 전체 반주 프레임 개수에 대한 false alarm 프레임의 개수의 비이다. 그리고 error ratio는 전체 프레임 개수에 대한 miss 혹은 false alarm 프레임 개수의 비를 말한다.

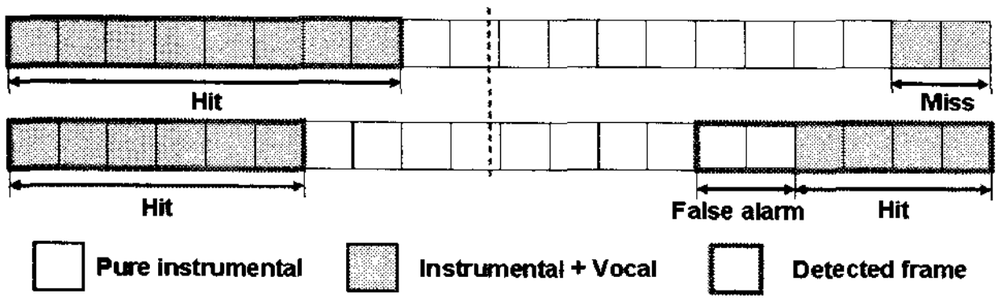


그림 5. Hit, miss, 그리고 false alarm 판별의 예
Fig. 5. Examples of hit, miss, and false alarm decisions

표 1에서는 식 (3)의 $B_l^{MR}(i)$ 의 크기만으로 보컬 영역과 반주 영역으로 분류하는 기본적인 TICFT 기법을 이용하였을 때 보컬/반주 영역 분류 결과를 제시하였다. 그리고 표 2에서는 III.1절에서 설명한 제안한 기법을 이용하였을 때의 보컬/반주 영역 분류 결과를 제시하였다. 표 1과 표 2를 비교하면, 제안한 기법은 기본적인 TICFT 기법에 비해 성능을 개선시킴을 알 수 있다. 보다 구체적으로 살펴보면, "유미"의 경우 miss ratio와 false alarm ratio를 모두 크게감소시켜, error ratio가 12.5%에서 5.8%로 향상되었다. "이

표 1. 기본적인 TICFT 기법을 이용한 보컬과 반주 영역 분류 결과
Table 1. Results of vocal and instrumental region classification using the basic TICFT method

	Hit ratio	Miss ratio	False alarm ratio	error ratio
유미	87.1 %	12.9 %	12.3 %	12.5 %
이승철	78.9 %	21.1 %	3.6 %	13.6 %
보아	53.2 %	46.8 %	5.6 %	44.1 %
미상	53.3 %	46.7 %	6.6 %	37.6 %

표 2. 제안한 기법을 이용한 보컬과 반주 영역 분류 결과
Table 2. Results of vocal and instrumental region classification using the proposed method

	Hit ratio	Miss ratio	False alarm ratio	error ratio
유미	97.6 %	2.4 %	7.8 %	5.8 %
이승철	99.6 %	0.4 %	27.6 %	9.6 %
보아	80.4 %	19.6 %	0.0 %	16.5 %
미상	83.8 %	16.2 %	8.5 %	14.8 %

승철"의 경우 false alarm ratio는 오히려 커졌지만, 전체적인 error ratio는 13.6%에서 9.6%로 개선된 성능을 보였다. "이승철"에서 false alarm ratio가 높아진 것은 곡의 건반 악기와 타악기의 조합이 보컬의 주파수 영역에 크게 영향을 주었기 때문으로 판단된다. 한편, "보아"와 "미상"의 곡에서는 miss ratio를 크게 낮추어 전체적인 error ratio를 크게 개선시켰다. 전체적으로 제안한 방법은 문턱값을 식 (5)와 (6)을 통해 국부적인 특성을 반영할 수 있도록 함으로써 영역 판별의 정확도를 높였음을 알 수 있다. 따라서 여기서 제시한 보컬/반주 영역 판별 기법을 이용함으로써 보컬 제거 기법의 성능을 개선시킬 수 있을 것으로 기대된다.

2. 보컬 제거 실험

그림 6-8에서는 "유미"에서 보컬 제거 과정을 단계별로 보여준다. 그림 6은 "유미" 원곡의 스펙트로그램이고, 그림 7은 IV.1절에서 얻은 보컬 영역에 대해서 보컬 성분만의 스펙트로그램이다. 그림 7을 참조하면, 보컬 영역에 대해서만 250~9000Hz의 주파수 성분이 나타남을 알 수 있다. 그림 6의 원곡에서 그림 7의 보컬 성분을 빼면, 그림 8과 같이

그림 6. "유미"의 스펙트로그램.
Fig. 6. The spectrogram of the "Yumi".

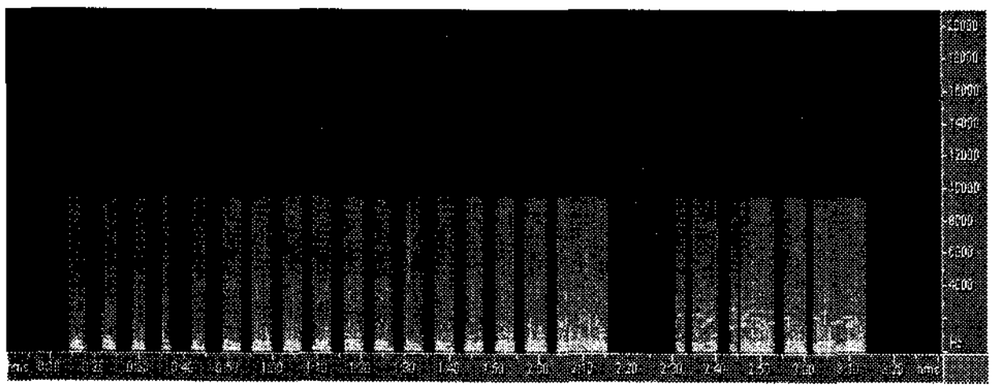


그림 7. "유미"의 추출된 보컬 성분 스펙트로그램.
Fig. 7. The spectrogram of the extracted vocals of the "Yumi."

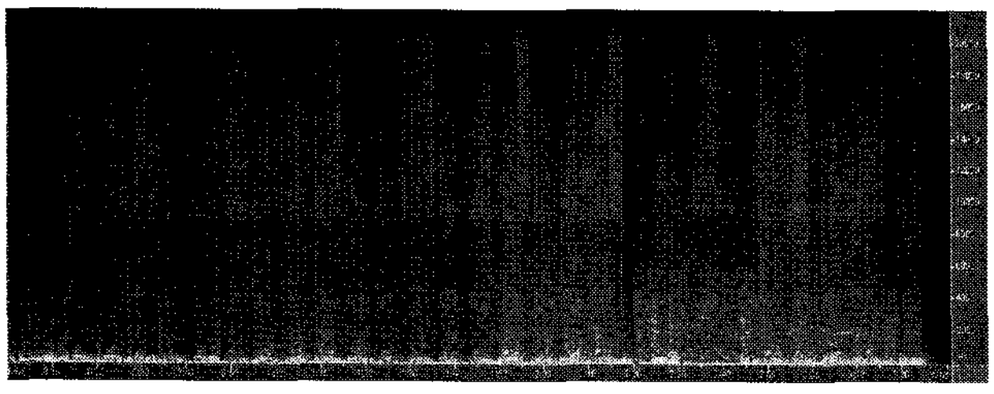


그림 8. "유미"에서 반주 성분의 스펙트로그램.
Fig. 8. The spectrogram of the instrumentals of the "Yumi."

반주 성분만 출력된다. 그림 8에 제시한 바와 같이 제안한 기법에서는 보컬 영역으로 판별된 프레임에 대해서 보컬 에너지가 집중된 대역만을 처리하였다는 것을 알 수 있다. 즉, 제안한 기법을 이용함으로써 보컬 성분만을 효과적으로 제거하고, 나머지 영역 및 주파수 성분은 그대로 출력함으로써 반주 성분을 충실히 출력할 수 있음을 보였다.

종합적으로 보면, 제안한 보컬 제거 기법은 Adobe 사의 audition 혹은 MS windows XP Media Center Edition (MCE)의 보컬 제거기능 등과 마찬가지로 크게 만족스러운 결과를 보여주지 않는다. MCE와 Audition의 성능은 곡에 따라 차이가 있지만, 경우에 따라 볼륨이 줄어든 것 같이 출력되는 등 전체적으로 비슷한 것으로 판단된다. 제안한 기법은 보컬 영역에서는 MCE와 Audition과 비슷한 성능을 보이며, 반주 영역에서는 원래의 음악이 그대로 재생된다. 그러나 특히 false alarm, miss 등과 같이 보컬 영역 검색이 제대로 이루어지지 않은 곳에서는 그 성능이 낮아진다. 실험 결과를 살펴보면, 보컬 영역이 제대로 추출되지 않은 부분은 주로 보컬 영역의 가장자리부분이다. 따라서 리듬 추출부를 적절히 구현하여 리듬에 따라 프레임을 나누어 보컬 영역 검색을 수행하면 성능이 개선될 것으로 기대된다. 이를 포함하여 보컬 영역 검색의 정확도를 높인다면, 제안한 보컬 제거 기법의 성능을 좀 더 개선시킬 수 있을 것으로 판단된다.

V. 결 론

본 논문에서는 오디오 신호에서 보컬 성분을 분리하는 보컬 제거 기법을 다루었다. 기존의 방식과 달리, 제안한 기법에서는 오디오 신호를 세부 프레임으로 나누고 각각의 프레임에 대해 보컬 영역 검색을 수행하였다. 그런 다음 보컬 프레임에 대해서만 보컬 제거 기법을 적용하였다. 보컬 영역 검색을 보컬 제거에 효율적으로 적용하기 위해, TICFT 영역에서 부대역 에너지에 문턱값을 설정하여 1차적으로 보컬 프레임과 반주 프레임으로 분류하였다. 그런 다음, 음성의 특성을 고려하여 보컬 프레임 전후의 반주 프레임에 대해 분류된 결과를 수정하였다. 그리고 보컬 제거를 위해 대역 통과된 보컬 프레임에 대해 보컬 성분을 추출하고, 이를 원래의 음악에서 감산하였다. 실험 결과, 본 논문에서 제시한 보컬 영역 검색 기법은 기본적인 방법에 비해 error ratio를 크게 낮출 수 있음을 보였다. 그리고 주관적인 품질 측면에서 보컬 제거 기법의 성능을 살펴보았다. 본 논문에서 제시한 보컬 제거 기법은 기존의 방법들에 비해 보다 생동감 있는 음악을 출력한다. 그러나 보컬 영역검색의 정확도를 향상시키고 보컬 성분을 추출하는 방법에 대해 추가적으로 연구한다면, 좀 더 개선된 성능을 얻을 수 있을 것으로 기대한다.

참 고 문 헌

[1] D. Taylor, "Distributed multimedia for a song," IEEE Parallel & Distributed Technology: Systems & Applications, vol. 4, no. 4, pp. 8-9, Winter 1996.

[2] Chang-Huai You, Han-Wu Sun, "Multi-band adaptive filtering application on vocal mute," 6th International Conference on Signal Processing, vol. 2, pp. 1711-1714, Aug. 2002.

[3] http://www.adobe.com/product/audition.

[4] http://www.microsoft.com/korea/windowsxp/mediacenter.

[5] Y. Li, D.L. Wang, "Separation of singing voice from music accom-

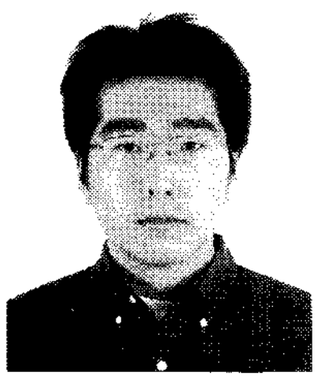
paniment for monaural recordings," IEEE Transactions on Audio, Speech and Language Processing, vol. 15, no. 4, pp. 1475-1487, May 2007.

[6] N.C. Maddage, Kongwah Wan, Changsheng Xu, and Ye Wang, "Singing voice detection using twice-iterated composite Fourier transform," IEEE International Conference on Multimedia and Expo, vol 2, pp. 1347~1350, June 2004.

[7] T.D. Rossing, F.R. Moore, P.A. Wheeler, The Science of Sound, Addison Wesley, 2002.

[8] J. Sundberg, The Science of the Singing Voice, Northern Illinois University Press, 1987.

저 자 소 개



김 유 승

- 2004년 : 서울시립대학교 전자전기컴퓨터공학부 학사 졸업
- 2006년 : 서울시립대학교 전자전기컴퓨터공학부 석사 졸업
- 2006년 ~ 현재 : 미디어코러스(주) 주임연구원
- 주관심분야 : 모바일 TV, MPEG system, 오디오 코덱

김 인 철

- 1985년 : 서울대학교 제어계측공학과 학사 졸업.
- 1987년 : 서울대학교 제어계측공학과 석사 졸업.
- 1992년 : 서울대학교 제어계측공학과 박사 졸업.
- 1992 ~ 1994년 : 대우전자(주) 영상연구소 선임연구원
- 1994 ~ 1999년 : 한성대학교 정보공학과 조교수
- 1999 ~ 현재 : 서울시립대학교 전자전기컴퓨터공학과 교수
- 주관심분야 : 신호처리, 음향처리, 실시간 신호처리, 오디오/비디오 코덱