특집논문 (Special Paper)
방송공학회논문지 제20권 제3호, 2015년 5월 (JBE Vol. 20, No. 3, May 2015)
http://dx.doi.org/10.5909/JBE.2015.20.3.408
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

에너지와 위상을 고려한 선택적 주파수 차감법을 이용한 보컬 분리

김 현 태[a], 박 장 식[b)‡]

Vocal Separation Using Selective Frequency Subtraction Considering with Energies and Phases

Hyuntae Kim[a] and Jangsik Park[b)‡]

요 약

최근 원음 반주기에 대한 관심이 증가됨에 따라 고가의 스튜디오 직접 녹음 방법 대신 보다 저렴한 방법을 시도하고 있다. 그 구체적인 방법으로는 가수의 음악 앨범에서 가수의 목소리만 제거하여 원음 반주 음원을 만드는 것이다. 본 논문에서는 보컬이 포함된 구간에서 스테레오로 녹음된 반주음악에서 보컬을 분리하는 시스템을 제안한다. 제안하는 시스템은 두 단계로 구성된다. 첫 단계는 보컬을 검출하는 단계이다. 이 단계에서는 MFCC를 가지고 SVM 방법을 이용하여 입력 신호를 보컬 부분과 비보컬 부분으로 분리한다. 두 번째 단계에서는 보컬 부분에 대해 각 주파수 빈별로 선택적 주파수 차감을 수행한다. 이 때 채널 신호의 주파수 빈별로 에너지 값뿐만 아니라 위상까지 고려하여 차감 여부를 판별한다. 제안하는 방법으로 보컬을 제거한 음악에 대한 청취 실험에서 상대적으로 높은 만족도를 보여준다.

Abstract

Recently, According to increasing interest to original sound Karaoke instrument, MIDI type karaoke manufacturer attempt to make more cheap method instead of original recoding method. The specific method is to make the original sound accompaniment to remove only the voice of the singer in the singer music album. In this paper, a system to separate vocal components from music accompaniment for stereo recordings were proposed. Proposed system consists of two stages. The first stage is a vocal detection. This stage classifies an input into vocal and non vocal portions by using SVM with MFCC. In the second stage, selective frequency subtractions were performed at each frequency bin in vocal portions. In this case, it is determined in consideration not only the energies for each frequency bin but also the phase of the each frequency bin at each channel signal. Listening test with removed vocal music from proposed system show relatively high satisfactory level.

Keyword : MFCC, SVM, Vocal Remover, Selective Frequency Subtraction, Inter-Channel Phase Difference

a) 동의대학교 멀티미디어공학과(Dept. of Multimedia Engineering, Dongeui University)
b) 경성대학교 전자공학과(Dept. of Electronics Engineering, Kyungsung University)
‡ Corresponding Author : 박장식(Jangsik Park)
E-mail: jsipark@ks.ac.kr
Tel: +82-51-663-4768
ORCID:http://orcid.org/0000-0003-1795-7631
※ 본 논문은 교육과학기술부에서 지원하는 한국연구재단(NRF) 기초원천기술개발사업(2014M3C1A1048865)의 연구결과입니다.
※ 본 논문은 BB21사업으로 지원한 결과입니다.
※ 이 논문의 연구결과 중 일부는 “IWAIT 2015”에서 발표한 바 있음.
· Manuscript received March 16, 2015; revised April 28, 2015; accepted May 4, 2015.

Ⅰ. 서 론

보컬이 포함된 노래에서 목소리 영역을 찾는 문제는 음성 특징으로 널리 사용되어진 전통적인 통계적인 접근법이 적용되어져 왔다. 예를 들면, GMM(Gaussian Mixture Model)[1-3], 신경망, 그리고 SVM 또는 HMM(Hidden Markov Model)이 사용되어져 왔다[4][5].

보컬 검출에 이어 보컬 분리는 두드러지는 음원의 음정을 추정하고 추정된 음정을 기반으로 해당 음원의 주파수 분포를 획득하여 마스크 또는 행렬 분해 기법을 활용하여 분리하는 방법이 진행되어 왔다[6][7]. 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.

본 논문에서는 보컬이 포함된 음악 구간에서 스테레오 음악 신호에서 보컬의 음상이 주로 센터에 위치한다는 사실에 기반한 주파수빈별 에너지 차감법과 채널 신호간 위상차 정보를 동시에 적용한 보컬 분리방법을 제안한다.

본 논문의 구성은 다음과 같다. 제 2장에서 본 논문에서 제안하는 보컬 분리 시스템에 대하여 설명하고 제 3장에서 실험 환경 및 결과에 대해 언급하며 제 4장에서 실험 결과를 기반으로 결론을 맺는다.

Ⅱ. 제안하는 보컬 분리 시스템

1. MFCC를 활용한 SVM 기반 보컬 검출

MFCC 특징 값들을 SVM을 이용하여 학습시킨다. SVM은 일반화 오차를 최소화할 수 있는 방향으로 학습을 수행하는 선형 분류기에서 비롯되었다. 그러나 선형 분리가 불가능한 경우, 고차원 매핑을 통해 해결할 수 있으나 계산량의 증가와 같은 부작용이 발생한다. 이러한 부작용을 해결하기 위해 제안된 방법이 커널 함수를 이용한 SVM 방법이다[8]. SVM 방법을 통해 학습과 분류를 수행하는 구체적인 절차는 아래와 같다.

(1) N개의 입출력 쌍으로 이루어진 학습데이터 집합 $\boldsymbol{X}=(\boldsymbol{x_i},y_i)_{i=1\cdots N}$을 준비하고 하이퍼 파라미터 c와 커널 함수 $k(\boldsymbol{x_i},\boldsymbol{x_j})$를 정의한다. 이 때 목표 출력값은 $y_i\in\{-1,1\}$ $(i=1,...,N)$을 만족하도록 정한다.

다음과 같은 과정을 통해 SVM을 학습한다.

(2) 학습데이터를 이용, 파라미터 추정을 위한 목적함수 $Q(\boldsymbol{\alpha})$를 정의한다.

$$Q(\boldsymbol{\alpha})=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j k(\boldsymbol{x_i},\boldsymbol{x_j}), \tag{1}$$

여기서 $\sum_{i=1}^{N}\alpha_i y_i=0,\ 0\le\alpha_i<c\ (i=1,...,N)$

주어진 조건을 만족하면서 $Q(\boldsymbol{\alpha})$를 최소화하는 추정치 $\hat{\alpha_i}$를 이차계획법(quadratic program)으로 찾는다.

$\hat{\alpha_i}\neq 0$ 이 되는 서포트벡터를 찾아 집합 $\boldsymbol{X_s}=\left\{\boldsymbol{x_i}\in\boldsymbol{X}|\hat{\alpha_i}\neq 0\right\}$ 를 생성한다.

$\hat{\alpha_i}$ 와 서포트벡터 이용하여 $\hat{\omega_o}$를 계산한다.

$$\hat{\omega_o}=\frac{1}{N_S}\sum_{\boldsymbol{x_i}\in\boldsymbol{X_S}}\left(y_i-\sum_{\boldsymbol{x_j}\in\boldsymbol{X_S}}\hat{\alpha}y_j\boldsymbol{x_j^T}\boldsymbol{x_i}\right) \tag{2}$$

이 때 N_S는 집합 $\boldsymbol{X_S}$의 원소의 수이다.

서포트벡터 집합 $\boldsymbol{X_S}=\left\{x_i\in\boldsymbol{X}|\hat{\alpha_i}\neq 0\right\}$와 파라미터 벡터 $\hat{\boldsymbol{\alpha}}$, 그리고 $\hat{\omega_o}$를 저장해 둔다.

(3) 새로운 데이터 $\boldsymbol{x}$가 주어지면, 저장해둔 서포트벡터와 파라미터를 이용하여 아래 함수로 분류를 수행한다.

$$f(\boldsymbol{x})=sign\left(\sum_{\boldsymbol{x_i}\in X_S}\hat{\alpha_i}y_i k(\boldsymbol{x_i},\boldsymbol{x})+\hat{\omega_o}\right) \tag{3}$$

2. 보컬 분리

일반적으로 보컬이 포함된 구간에서는 보컬의 목소리가 청취자의 귀에 들려야 하므로 배경 음악보다 에너지가 높다. 또한 보컬 성분은 음상의 정위에서 중앙 부근에 위치한

다. 따라서 좌, 우 스테레오 채널에 보컬 성분이 균일하게 분포하는 특성을 갖는다. 이러한 특성을 고려하여 보컬 구간에서의 좌, 우 채널간 차신호에는 보컬 성분이 제거되고 남은 상대적으로 낮은 에너지를 갖는 배경 음악 성분만 존재한다. 이러한 점을 이용하기 위해 좌, 우 스테레오 채널 신호와 차신호의 스펙트럼 상의 주파수 빈별 에너지를 비교하여 일정 기준값 이상 좌, 우 스테레오 채널 신호가 큰 경우 보컬이 포함된 주파수 빈으로 판별할 수 있다[9]. 그러나 록 음악이나 해비메탈 등 비트가 강한 음악은 보컬이 포함되지 않은 경우라도 에너지가 커 음상이 정면 부근에서 크게 벗어나지 않은 경우 보컬 성분으로 잘못 분류될 수 있다. 따라서 이 경우는 채널간 위상 차를 고려하면 상대적으로 채널 간 위상차가 크게 나타나 보컬로 분류되는 것을 방지할 수 있다. 그림 1에 이러한 부분을 표시하였다.

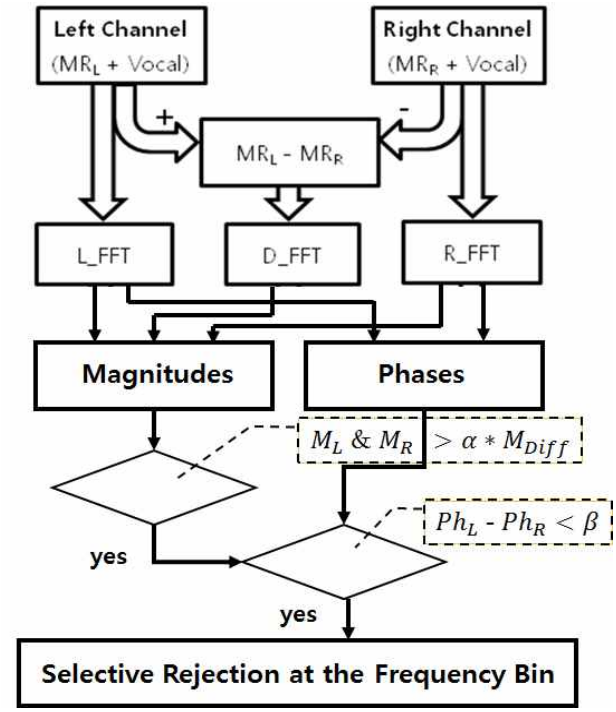

그림 1. 보컬 제거를 위한 상세 블록도
Fig. 1. Detailed block diagram for vocal removal

표 1은 그림 1의 블록도에 따라 시스템이 처리하는 절차를 나타낸 것이다.

표 1. 보컬 제거 절차
Table 1. Processing for vocal removal

Stage	Details
Step 1	Compute $MR_L - MR_R$ in time domain
Step 2	Transform each channel of the stereo signal in time domain into frequency domain by FFT
Step 3	Compute magnitudes of each channel and $MR_L - MR_R$ channel in frequency domain
Step 4	Implement spectral power comparison between each channel of the stereo signal and inter-channels difference
Step 5	Implement phase difference comparison between each channel of the stereo signal
Step 6	Reject selectively at each vocal frequency bin in stereo channel

Ⅲ. 실험 환경 및 결과

다양한 장르의 대중 음악을 모두 동일한 조건으로 실험하기위해 샘플주파수를 44100 ㎐로 고정하여 적용하였다. 또한 한 프레임 당 샘플 수는 16384로 두었고, 50%씩 오버랩하며 처리하였다. 주파수 해상도는 프레임당 샘플 수와 동일한 개수로 두었다. 한 프레임당 MFCC 계수는 정규화 에너지 파라미터 한 개를 포함하여 모두 13 개를 가진다. 또한 SVM에서 사용한 커널은 가우스 커널이다. 보컬을 제거하고 남은 배경 음악에 대한 음질 열화에 대한 평가이다. 평가는 MOS(mean opinion score) 테스트로 하였으며 한 곡 당 5점 만점으로 5단계로 나누어 평가하며, 다섯 가지 장르의 10개 음원에 대한 처리 결과를 가지고 10명의 청취자를 선정하여 테스트 전에 사전 교육을 통해 미리 단계별 음질 열화 정도를 비교 청취 후 실시하였다. 표 2는 청취 테스트에 사용한 음원을 나타낸다.

또한 보컬 구간에서 반주가 혼재하는 경우에 대한 위상값의 추이를 이해하기 위해 특정 곡(바이브의 “술이야 (Suliya)”)에서 보컬과 반주가 함께 들어있는 17번째 처리 블록에 대해 추출한 위상 값을 그림 2와 3에 나타내었다. 그림 2의 (a)는 왼쪽 채널이며, (b)는 오른쪽 채널, (c)는 좌, 우 차신호에 대한 것이며, (d)는 왼쪽 채널 위상값에서 오른

쪽 채널 위상값을 뺀 값을 나타낸다. 그림 3에 이를 보다 확대하여 나타내었다. 점선으로 표시된 타원이 보컬 성분에 해당하며, 보컬구간에 위상차가 줄어드는 것을 확인할 수 있다. 표 3에 언급한 청취 테스트 결과를 살펴보면 기존 결과 [9]와 비교하여 록 음악과 팝 음악에 속하는 곡에서 다소 청취 결과가 향상된 것을 확인할 수 있다.

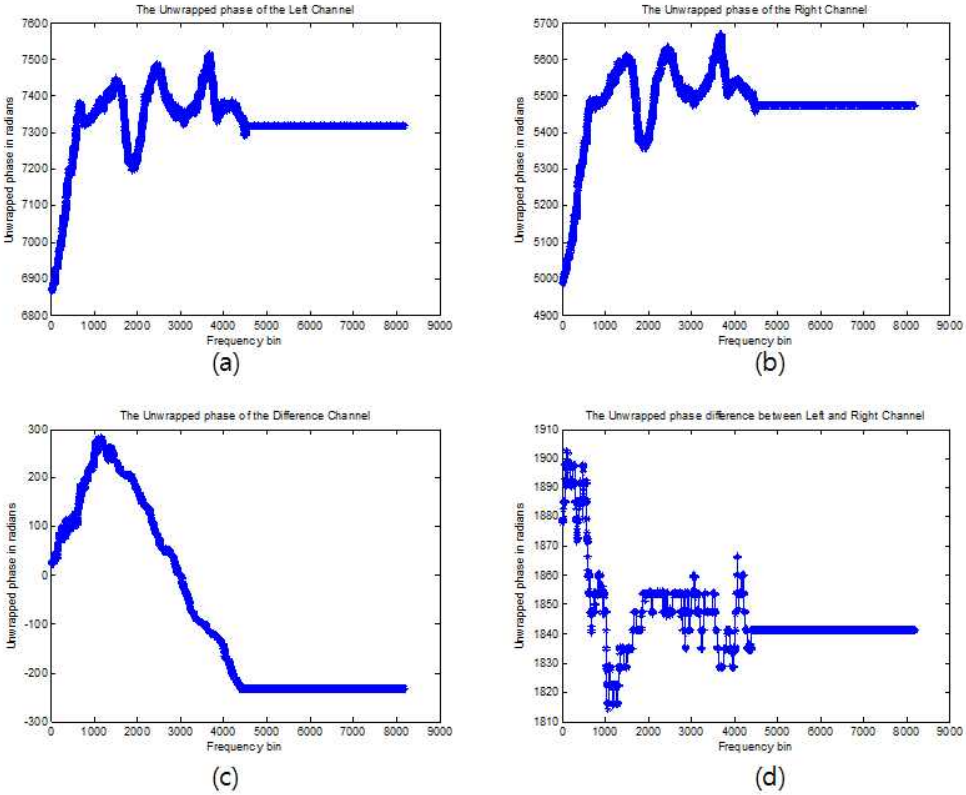

그림 2. 샘플 신호에 대한 위상 변화 예(보컬과 반주 혼재 구간)
Fig. 2. an example of phases variation for sample signal (vocal range)

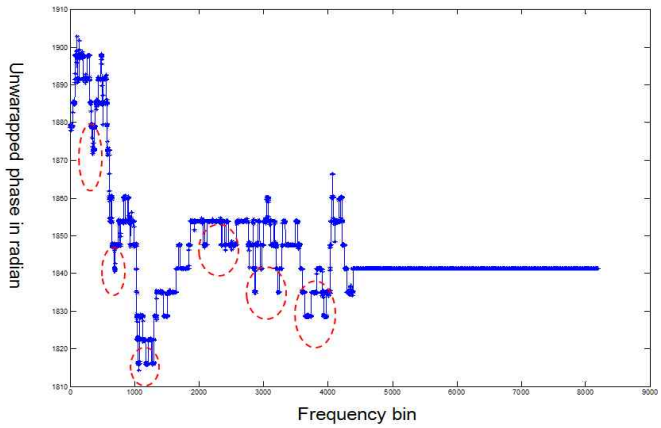

그림 3. 그림 2 (d) 확대 도면
Fig. 3. Enlarged plot in Fig. 2 (d)

표 2. 청취 테스트에 사용한 음원
Table 2. The music sources for listening test

Genres	Title	Singer
Ballad	Suliya(Korean, 'Alcohol is!')	Vibe
	Haneuleul Boa (Korean, Look at the sky!)	Kang, Chan
Trot	Eomeona (Korean, 'Goodness!'))	Jang, Yunjung
	Ichasun Dari (Korean, 'Two lane Bridge')	Cha, Taehyun
Rock	Sarang (Korean, 'Love')	Bu Hwal
	This Is How We Stand	Mirva
Pop	Love Song	Sara Bareilles
	Before I Say Goodbye	Lauren Piper
vocal music	Bimok (Korean, tombstone made by wood)	John Park
	Hyangsu (Korean, homesickness)	Lee Dongwon, Shin Dong-ho

표 3. 청취 테스트 결과(평균 점수)
Table 3. The results for listering test(average score)

listener	Genres										average /listener
	Ballad		Trot		Rock		Pop		vocal music		
	1	2	1	2	1	2	1	2	1	2	
A	4	4	4	3	3	3	3	3	4	4	3.50
B	4	4	4	4	3	3	3	4	4	4	3.70
C	4	3	4	3	3	3	3	3	4	4	3.40
D	4	3	4	4	3	3	3	3	4	4	3.50
E	4	4	4	4	3	3	3	3	3	4	3.50
F	4	4	4	4	3	3	4	4	4	4	3.80
G	4	4	4	4	3	4	3	3	4	4	3.70
H	4	3	4	4	3	3	3	3	4	4	3.50
I	4	4	4	4	3	3	3	3	4	4	3.60
J	4	4	4	4	3	3	3	3	4	4	3.60
K	4	4	4	4	3	3	3	3	4	3	3.50
	total average										3.57

Ⅳ. 결 론

본 논문에서는 스테레오 음악 신호에서 보컬의 음상이 주로 센터에 위치한다는 사실에 기반한 주파수빈별 에너지 차감법을 적용한 보컬 분리방법을 개선하여 좌, 우 스테레오 채널신호간 위상차 정보를 추가하여 록(rock) 이나 팝(pop) 등 반주 음악의 에너지가 큰 장르에서 보컬 제거 성능을 향상하는 시스템을 제안하였다. 청취테스트를 통한 보컬 분리 성능이 기존에 제안한 방법[9]에 비해 전체 평균 3.57 점으로 약 0.05점 정도 향상되었으며, 주로 록(rock) 이나 팝(pop)에서 향상된 것으로 판단된다. 따라서 제안하는 방법이 보컬 분리 성능 향상에 기여할 수 있는 것을 확인하였다.

참 고 문 헌 (References)

[1] H. Kim, G. Lee, J. park, and Y. Yu, "Vehicle Detection in Tunnel using Gaussian Mixture Model and Mathematical Morphological Processing," J. of the Korea Institute of Electronic Communication Science, vol. 7, no. 5, 2012, pp. 967-974.

[2] K. Park and H. Kim, "A Study for Video-based Vehicle Surveillance on Outdoor Road," J. of the Korea Institute of Electronic Communication Science, vol. 8, no. 11, 2013, pp. 1647-1653.

[3] H. Kim and J. Park, "Smoke Detection in Outdoor Using Its Statistical Characteristics," J. of the Korea Institute of Electronic Communication Science, vol. 9, no. 2, 2014, pp. 149-154.

[4] T. Leung, C. Ngo, and R.W.H. Lau, "Ica-fx features for classification of singing voice and instrumental sound," in Proc. International Conference on Pattern Recognition, Cambridge, UK, 2004, vol. 2.

[5] A. Berenzweig and D.P.W. Ellis, "Locating singing voice segments within music signals," in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'2001), New York, USA, October, 2001.

[6] T. Virtanen, A. Mesaros, and M. Ryynänen, "Combining Pitch-Based Inference and Non-Negative Spectrogram Factorization in Separating Vocals from Polyphonic Music," Proc. Statistical and Perceptual Audition, Brisbane, Australia, September 2008.

[7] J.-L. Durrieu, A. Ozerov, C. Févotte, G. Richard and B. David, "Main instrument separation from stereophonic audio signals using a source/filter model," 17th European Signal Processing Conference (EUSIPCO 2009) Glasgow, Scotland, August 2009.

[8] Hae Y. Park, Kwan Y. Lee, "Pattern and Machine Learning from Fundamental to Applications, Ihan Press, Goyang, South Korea, 2011.

[9] H. Kim, "Vocal Separation in Music Using SVM and Selective Frequency Subtraction" J. of the Korea Institute of Electronic Communication Science, vol. 10, no. 1, 2015, pp. 1-6.

저 자 소 개

김 현 태

- 1989년 : 부산대학교 전자공학과 졸업(공학사)
- 1995년 : 부산대학교 대학원 전자공학과 졸업(공학석사)
- 2000년 : 부산대학교 대학원 전자공학과 졸업(공학박사)
- 2002년 ~ 현재 : 동의대학교 멀티미디어공학과 교수
- ORCID : http://orcid.org/0000-0001-9608-0743
- 주관심분야 : 영상 및 음향신호처리, 적응신호처리

박 장 식

- 1992년 : 부산대학교 전자공학과 졸업(공학사)
- 1994년 : 부산대학교 대학원 전자공학과 졸업(공학석사)
- 1999년 : 부산대학교 대학원 전자공학과 졸업(공학박사)
- 1997년 ~ 2011년 : 동의과학대학 디지털전자과 교수
- 2011년 ~ 현재 : 경성대학교 전자공학과 교수
- ORCID : http://orcid.org/0000-0003-1794-7631
- 주관심분야 : 적응신호처리, 영상 및 음향신호처리, 임베디드시스템