Regular paper http://dx.doi.org/10.13067/JKIECS.2015.10.1.1

SVM과 선택적 주파수 차감법을 이용한 음악에서의 보컬 분리

김현태*

Vocal Separation in Music Using SVM and Selective Frequency Subtraction

Hyun-Tae Kim*

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 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. 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. Listening test with removed vocal music from proposed system show relatively high satisfactory level.

키워드

MFCC, SVM, Vocal Remover, Selective Frequency Subtraction
MFCC, SVM, 보컬 제거, 선택적 주파수 차감

Ⅰ. 서 론

최근 원음노래반주기 수요가 증가함에 따라 노래반주기 시장에 미디기반 노래반주기 대신 원음노래반주기가 보급되고 있다. 그러나 원음 노래 반주음악을 제작하는 데 많은 비용이 소요되는 문제로 보급이 지연되고 있어 보다 저비용으로 원음 반주음악을 제작하는 방법이 요구되고 있다. 그 중 한 방법으로 제시되

* 교신저자(corresponding author) : 동의대학교 멀티미디어공학과(htaekim@deu.ac.kr)
접수일자 : 2014. 10. 09 심사(수정)일자 : 2014. 12. 15 게재확정일자 : 2015. 01. 12

는 것은 가수의 원음 음반에서 가수의 보컬만 제거하여 원음 반주음악으로 사용하는 것이다. 이를 위해서는 우선적으로 음악 속에 보컬이 존재하는 지를 자동으로 판별하는 보컬 검출 기술이 필요하다. 보컬 영역의 자동 검출은 가수 자동 식별, 보컬 분리 등을 포함한 다양한 응용분야에 필수적인 단계이다[1-2].

보컬이 포함된 노래에서 목소리 영역을 찾는 문제는 음성 특징으로 널리 사용되어진 전통적인 통계적인 접근법이 적용되어져 왔다. 예를 들면, GMM(Gaussian Mixture Model)[3-5], 신경망, 그리고 SVM 또는 HMM(Hidden Markov Model)이 사용되어져 왔다[6-7].

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[8-9]. 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 인간의 청각 특성을 고려한 MFCC (Mel-Frequency Cepstral Coefficients) 관련 특징 값들에 대한 보컬 영역과 비보컬 영역의 차이를 SVM (Support Vector Machine) 방법을 통해 훈련하고 판별하는 보컬 검출 부분과 스테레오 음악 신호에서 보컬의 음상이 주로 센터에 위치한다는 사실에 기반한 주파수빈별 에너지 차감법을 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

2.1 MFCC를 활용한 SVM 기반 보컬 검출

스펙트럼 특성을 뽑아내는 가장 인기있고 우세한 방법은 MFCC 이다. MFCC는 비선형적인 Mel스케일의 주파수 도메인에서 로그파워스펙트럼에 코사인변환(cosine transform)을 취함으로써 얻을 수 있다. MFCC와 일반적인 켑스트럼의 차이는 일반적인 켑스트럼의 경우 주파수 밴드가 균등하게 나누어져 있는 반면 MFCC의 경우 주파수 밴드가 Mel-scale에서 균등하게 나누어진다는 것이다.

MFCC는 일반적으로 다음의 과정을 통해 구할 수 있다.

① 단구간 음성에 Fourier Transform을 취한다.
② 위 값들에서 Mel-scale의 필터뱅크를 이용해 파워스펙트럼을 구한다.
③ 각 Mel-scale의 파워에 로그를 취한다.
④ 위 값에 discrete cosine transform 을 취하면 MFCCs 값이 나온다.

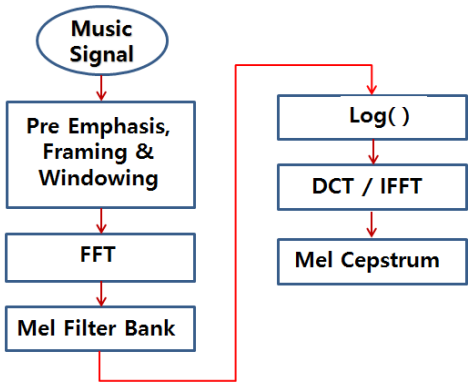

그림 1. MFCC 추출 과정
Fig. 1 Process for extracting MFCC

이렇게 추출한 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[10]. SVM 방법을 통해 학습과 분류를 수행하는 구체적인 절차는 아래와 같다.

1. N개의 입출력 쌍으로 이루어진 학습데이터 집합 $\boldsymbol{X}=(\boldsymbol{x}_i,y_i)\ (i=1,...,N)$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.

2. 다음과 같은 과정을 통해 SVM을 학습한다.

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

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

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

나. 주어진 조건을 만족하면서 $Q(\boldsymbol{\alpha})$를 최소화하는 추정치 $\hat{\alpha_i}$를 이차계획법으로 찾는다.

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

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

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

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

라. 서포트벡터 집합 $\boldsymbol{X_S} = \{x_i \in \boldsymbol{X} | \hat{\alpha_i} \neq 0\}$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.

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

2.2 보컬 분리

보컬로 분류된 영역에서 보컬을 제거하기 위해, 스테레오 신호의 경우 대부분의 보컬이 센터 중앙에 정위한다는 사실에 착안한다. 따라서 보컬이 포함된 영역인 경우, 좌우 스테레오 신호간 차신호에는 보컬 성분이 제거되고 남은 배경 음악만 남게되어 좌우 스테레오 신호에 비해 일정 비율 보다 낮은 양의 에너지를 가질 것이다. 이를 프레임별 및 주파수빈별로 비교하여 동일 조건인 주파수는 보컬 성분 주파수로 판단하여 제거하고 그렇지 않은 주파수는 그대로 살린다. 이러한 처리를 통해 보컬을 제거한다.

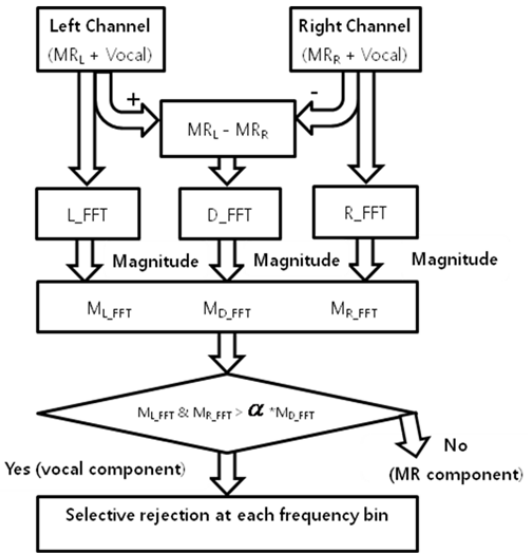

그림 2. 보컬 제거를 위한 상세 블록도
Fig. 2 Detailed block diagram for vocal removal

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	Reject selectively at each vocal frequency bin in stereo channel

보컬 제거 절차는 그림 2과 표 1에 보다 상세히 나타내었다.

III. 실험 환경 및 결과

다양한 장르의 대중 음악을 모두 동일한 조건으로 실험하기위해 샘플주파수를 16000 Hz로 고정하여 그보다 상위의 샘플주파수를 가진 데이터는 이에 맞게

리샘플링하여 적용하였다. 또한 한 프레임 당 샘플 수는 400으로 두었고, 한 프레임당 MFCC 계수는 정규화 에너지 파라미터 한 개를 포함하여 모두 13 개를 가진다. 또한 SVM에서 사용한 커널은 가우스 커널이다.

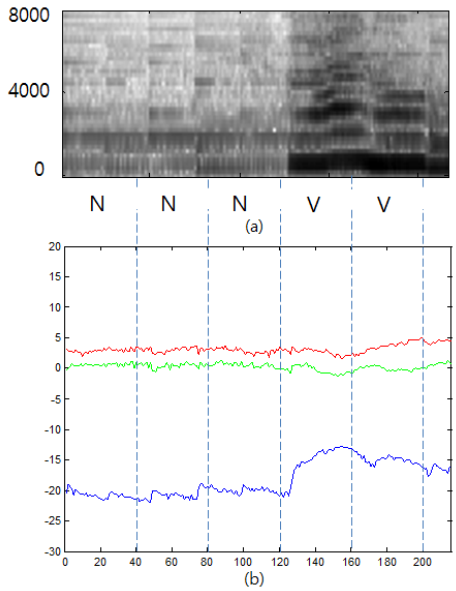

그림 3. 보컬과 비보컬 영역에 대한 MFCC 계수 값 비교
(a) 스펙트로그램
(b) MFCC 계수 중 처음 3개 값 변화
Fig. 3 Detailed block diagram for vocal removal
(a) spectrogram
(b) Variation for the first three coefficients of MFCC

그림 3의 (a)는 특정 음악에 대한 스펙트로그램으로 보컬이 포함된 영역은 'V'로 순수 악기 연주 영역은 'N'으로 표시하였으며 1초 간격으로 나누어 표시한 것이다. 그 이유는 세그먼트 간격을 더 세분화하면 세그먼트 단위로 청취 시 보컬과 비보컬을 인지하기 어려워지기 때문이다. 그리고 그림 3의 (b)는 동일 음악에 대한 MFCC 계수 중 처음 3개를 표시한 것이며, 가장 아래에 있는 곡선이 첫 번째 계수이며, 맨 위에 위치한 곡선이 두 번째 계수를 나타낸다.

그림 3의 (b)에 나타난 바와 같이 보컬 영역에서 MFCC 계수가 상당히 변화되는 것을 볼 수 있다.

다양한 장르의 음악에 대한 실험 결과는 표 2에 나타내었다. 보컬과 비보컬 영역은 1초 간격의 세그먼트로 구분하여 보컬 영역을 보컬이라고 판별한 세그먼트 수 및 분류 성공률을 가장 우측에 표시하였다.

두 번째 단계는 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 3은 청취 테스트에 사용한 음원을 나타낸다. 보컬 분리 성능은 표 4에 나타내었으며, 10명의 청취자에 대한 청취 테스트 결과 전체 평균 3.527 점으로 나타났다.

표 2. 보컬 영역 검출 결과
Table 2. The results for vocal region detection

Genres	Title	Singer	Length(time)	Vocal Segment	Detecting Segment(number/rate(%))
Ballad	Suliya(Korean, 'Alcohol is!')	Vibe	0:22	14	12/85.71
Trot	Eomeona(Korean, 'Goodness!'))	Yunjung Jang	0:27	24	20/83.33
Rock	Sarang(Korean, 'Love')	Bu Hwal	0:29	29	19/65.52
	This Is How We Stand	Mirva	1:03	52	34/65.38
Pop	Love Song	Sara Bareilles	0:25	21	16/76.19
	Before I Say Goodbye	Lauren Piper	0:40	21	17/80.95
	Average				76.18

표 3. 청취 테스트에 사용한 음원
Table 3. 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 4. 청취 테스트 결과(평균 점수)
Table 4. 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	2	3	3	4	4	3.4
B	4	4	4	4	3	3	3	4	4	4	3.7
C	4	3	4	3	3	3	3	3	4	4	3.4
D	4	3	4	4	3	3	3	3	4	4	3.5
E	4	4	4	4	3	3	3	3	3	4	3.5
F	4	4	4	4	3	3	3	3	4	4	3.6
G	4	4	4	4	3	3	3	3	4	4	3.6
H	4	3	4	4	3	3	3	3	4	4	3.5
I	4	4	4	4	2	3	3	3	4	4	3.5
J	4	4	4	4	3	3	3	3	4	4	3.6
K	4	4	4	4	3	3	3	3	4	3	3.5
	total average										3.527

Ⅳ. 결 론

본 논문에서는 인간의 청각 특성을 고려한 MFCC (Mel-Frequency Cepstral Coefficients) 관련 특징 값들에 대한 보컬 영역과 비보컬 영역의 차이를 SVM(Support Vector Machine) 방법을 통해 훈련하고 판별하는 보컬 검출 부분과 스테레오 음악 신호에서 보컬의 음상이 주로 센터에 위치한다는 사실에 기반한 주파수빈별 에너지 차감법을 적용한 보컬 분리 방법을 제안하였다. 보컬 검출 성능은 직접 청취하여 세그먼트별 판별한 내용과 비교한 결과 76.18% 일치하여 검출 기준으로 비교적 좋은 성능을 보이나 고려할 특징수를 늘려 일치율을 보다 높일 필요가 있다. 또한 보컬 분리 성능은 청취 테스트 결과 전체 평균 3.527 점을 보이나 주로 록과 팝 음악에서 분리 성능이 다소 줄어드는 경향이 있어 상대적으로 보컬과 반주 음악간 에너지 차이가 적은 경우 분리 성능이 저하되는 것으로 판단된다.

감사의 글

이 논문은 2013학년도 동의대학교 교내연구비에 의해 연구되었음(2013AA169).

References

[1] W. Tsai and H. Wang, "Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, issue 1, 2006, pp. 330-341.

[2] S. Vembu and S. Baumann, "Separation of vocals from polyphonic audio recordings," In *Proc. Int. Society for Music Information Retrieval*, London, UK, Sept., 2005.

[3] H. Kim, G. Lee, J. park, and Y. Yu, "Vehicle Detection in Tunnel using Gaussian Mixture Model and Mathematical Morphological Proc-

essing," *J. of the Korea Institute of Electronic Communication Science*, vol. 7, no. 5, 2012, pp. 967-974.

[4] 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.

[5] 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.

[6] T. Leung, C. Ngo, and R. W. H. Lau, "Ica-fx features for classification of singing voice and instrumental sound," In *Proc. Int. Conf. on Pattern Recognition,* Cambridge, UK, vol. 2, Aug. 2004.

[7] A. Berenzweig and D. P. W. Ellis, "Locating singing voice segments within music signals," In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'2001),* New York, NY, Oct. 2001.

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

[9] 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 Conf. (EUSIPCO 2009)* Glasgow, Scotland, Aug. 2009.

[10] H. Park and K. Lee, "*Pattern and Machine Learning from Fundamental to Applications*, Goyang, Korea : Ihan Press, 2011.

저자 소개

김현태(Hyun-Tae Kim)

1989년 부산대학교 전자공학과 졸업(공학사)
1995년 부산대학교 대학원 전자공학과 졸업(공학석사)
2000년 부산대학교 대학원 전자공학과 졸업(공학박사)
2002년～현재 동의대학교 멀티미디어공학과 교수
※ 관심분야 : 영상 및 음향신호처리, 적응신호처리