확률적 은닉 성분 분석 및 음향 기술자를 사용한 내용 기반 음악 요소 검색 방법

한병준*, 이교구**, 노승민*, 황인준*
*고려대학교 전기전자전파공학부
**서울대학교 융합과학기술대학원 디지털정보융합학과
*{hbj1147, smrho, ehwang04}@korea.ac.kr, **kglee@snu.ac.kr

A Scheme for Content-based Music Element Retrieval Using Probabilistic Latent Component Analysis and Acoustic Descriptor

Byeong-jun Han*, Kyogu Lee**, Seungmin Rho*, Eenjun Hwang*
*School of Electrical Engineering, Korea University
**Graduate School of Convergence Science and Technology, Seoul National University

요 약

지금까지 음악 정보 검색을 위한 다양한 내용 기반 음악 검색 및 비교 방법이 제안되었다. 그런데, 기존 연구들은 질의 방식 및 검색 카테고리가 변화함에 따라 상이한 방법을 제시하고 있어 음악 검색 방법을 통합하는 데에 한계가 있다. 이러한 문제를 해결하기 위해, 본고에서는 내용 기반 음악 검색의 일반화를 위한 내용 기반 음악 요소 검색(CBMER) 방법을 제안하였다. 제안 방법에서는 확률적 은닉 성분 분석(PLCA)을 사용하여 음원을 분해하고, 각 분해 요소로부터 오디오 특성을 추출하였다. 제안 방법을 사용하여 다양한 질의 방식 및 검색 카테고리로 내용 기반 음악 요소 검색이 가능함을 보이기 위해, 남성/여성의 목소리로부터 질의를 생성하여 목소리 성별에 따른 음악을 검색하는 실험을 수행하고 그 결과를 분석하였다.

1. 서론

최근 네트워크 대역폭 향상 및 MP3 플레이어 등과 같은 휴대용 장치의 보급으로 음악 콘텐츠에 대한 접근성이 크게 향상되었다. 이에, 최근 음악을 분석하여 다양한 정보 처리에 활용하는 방법에 대한 연구 분야인 음악 정보 검색(MIR: Music information retrieval) 분야가 급성장하였다. 또한, 다양한 음악 정보 처리 방법들에 대한 평가[1]가 이루어지고 있다.

음악 정보 검색 연구 분야에서는 다양한 검색 카테고리에 따른 검색 방법이 제시되고 있다. 한편, 질의 방식 또는 검색 카테고리별 색인 작성 및 검색 방법은 질의 방식과 검색 카테고리의 조합에 따라 상이하다. 이러한 상황하에서, 제시된 모든 방법들을 하나의 종합 검색 시스템으로 통합하기 위해서는 각 조합별 검색 방법의 최적화 및 통합을 위한 거대하고 복잡한 작업이 필요하다. 이로 인해 전체 음악 정보 처리 시스템이 비대해져 처리 효율이 낮아질 수 있는 한계가 있다.

본 연구에서는 전술한 문제와 한계점을 해결하고 내용 기반 음악 검색 방법을 일반화하기 위한 내용 기반 음악 요소 검색(CBMER: Content-based music element retrieval) 방법을 제안한다. 내용 기반 음악 요소 검색은 다양한 질의 방식과 검색 카테고리에 대한 내용 기반 음악 검색 방법을 보다 일반화하므로 다양한 질의 방식 및 검색 카테고리에 대응할 수 있다.

제안하는 방법은 확률적 은닉 성분 분석(PLCA)[2]을 사용하여 노래 음원을 주파수 및 시계열 성분으로 분해하고, 이들로부터 다양한 음향 기술자 중 주파수 특성을 추출하여 색인화한다. 제안 방법의 효율성 및 효과성을 평가하기 위한 실험을 수행하고 그 결과를 분석한다.

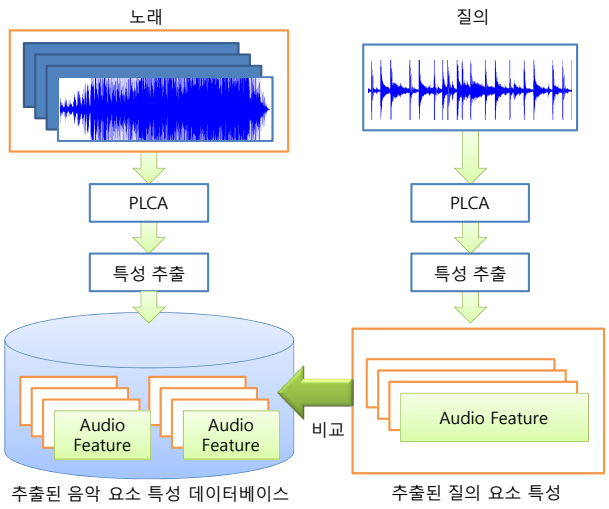
본고는 연구의 동기와 기존 연구의 한계점, 그리고 제안 방법을 요약한 본 장을 비롯, 총 5 장으로 구성된다. 2 장에서는 내용 기반 음악 검색 및 음악 요소 추출과 활용에 관련된 기존 연구와 그 한계점을 알아본다. 3 장에서는 제안하는 내용 기반 음악 요소 검색 방법을 상세히 설명한다. 제안 방법에 대한 실험 결과 및 분석은 4 장에서 이루어진다. 마지막으로 5 장에서는 제안 방법에 대한 결론을 제시한다.

2. 관련 연구

이 장에서는 제안하는 내용 기반 음악 요소 검색을 설명하기에 앞서, 기존의 내용 기반 음악 검색 관련 연구 및 음악 및 오디오 요소를 분석하는 다양한 접근 방법을 조사한다.

2.1. 내용 기반 음악 검색

내용 기반 음악 검색은 전문가가 수동으로 음악의 내용을 분석하는 것이 아닌, 자동화된 정보 처리 시



(그림 1) 내용 기반 음악 요소 검색 방법

스템을 사용하여 음악의 내용을 추출하여 검색에 활용하기 위한 방법이다. 지금까지 허밍[3] 및 태핑(tapping)[4]과 같은 다양한 질의 방식과 음악 감정 및 온톨로지[5], 무드 인식[6], 그리고 코드 및 장르 분류[7]와 같은 다양한 검색 카테고리를 위한 내용 기반 음악 검색 방법이 제안되어 있다. 그러나, 이런 다양한 질의와 검색 카테고리를 통합하는 하나의 시스템이 존재하지 않으며, 기존의 다양한 방법론을 하나의 시스템으로 통합하는 데에는 방법론이 비대해지는 등의 한계가 있다.

2.2. 음악 요소 추출 및 활용

노래에서 음악 요소를 추출하기 위해서는 노래를 음원 분리(source separation)하여 분석해야 할 필요가 있다. 이를 위해, Smaragdis *et al.*[2]은 오디오 모델링을 위한 확률적 은닉 성분 분석(PLCA)의 이론, 그리고 특성 추출, 음원 인식, 음원 분리, 노이즈 제거 등의 다양한 응용 사례를 제시하였다. 한병준 등[8][9]은 음악 요소 중 타악기 성분을 추출하기 위해 PLCA 를 사용하여 타악기의 onset 을 검출하고 음원을 분리하는 연구를 진행하였다. 한편, N. Cho *et al.*[10]은 음원을 Gabor atom 으로 분해하고 이로부터 검색 사전을 구축하였다.

3. 내용 기반 음악 요소 검색

이 장에서는 그림 1 과 같이 제안하는 내용 기반 음악 요소 검색 방법에 대해 설명한다. 제안하는 검색 방법은 음악 요소 분해, 음향 기술자를 이용한 특성 추출, 그리고 검색 방법으로 나뉜다.

3.1. 음악 요소 분해

확률적 은닉 성분 분석(PLCA: Probabilistic latent component analysis) [2]은 $\mathbf{x}=\{x_1,x_2,\cdots,x_n\}$에 대한 확률 밀도 함수 $P(\mathbf{x})$ 를 n 차원의 N 개 은닉 성분 $\mathbf{z}=\{z_1,z_2,\cdots,z_N\}$에 대한 다음과 같은 확률 분포 함수로 모델링하는 방법이다:

$$P(\mathbf{x})=\sum_{m=1}^{N}P(z_m)\prod_{k=1}^{n}P(x_k|z_m)$$

PLCA 모델링에는 다양한 메타 알고리즘이 적용 가능하다. EM 알고리즘의 경우[2], E-step 에서는 은닉 변수 z에 대한 기대치 R 을 다음과 같이 모델링한다:

$$R(\mathbf{x},z_m)=\frac{P(z_m)\prod_{j=1}^{n}P(x_j|z_m)}{\sum_{k=1}^{N}P(z_k)\prod_{j=1}^{n}P(x_j|z_k)}$$

M-step 에서는 다음으로 은닉 성분 분포 $P(\mathbf{z})$ 및 은닉 성분별 확률 분포 함수 $P(x_k|z_m)$를 최대화한다:

$$P(z_m)\leftarrow\int P(\mathbf{x})R(\mathbf{x},z_m)\mathrm{d}\mathbf{x}$$

$$P(x_k|z_m)\leftarrow\frac{\int\cdots\int P(\mathbf{x})R(\mathbf{x},z_m)dx_l,\forall l\neq k}{P(z_m)}$$

한편, 음원 분해를 위하여 주파수 f와 시간 t에 대한 $\mathbf{X}_2=\{f,t\}$에 대해 스펙트로그램 에너지 확률 밀도 함수 $\mathrm{P}(\mathbf{X}_2)$를 2 차원 PLCA 로 모델링하면,

$$P(\mathbf{X}_2)=\sum_{m=1}^{N}P(z_m)P(f|z_m)P(t|z_m)$$

과 같다.

이러한 2 차원 PLCA 모델에서 EM 알고리즘은 일반차원 PLCA 에 적용하는 EM 알고리즘으로부터 유도할 수 있다[8].

한편, 2차원 PLCA 모델에서 $P(f|z_m)$ 및 $P(t|z_m)$은 각각 주파수 및 시계열 성분으로, $P(f|z_m)$은 음원의 m번째 주파수 성분 패턴을, $P(t|z_m)$은 시간 흐름에 따른 m번째 주파수 성분의 출현 비중을 나타낸다.

3.2. 색인화를 위한 스펙트럼 특성 추출

PLCA 에 의해 분해된 N개의 주파수 성분은 시계열 성분 및 은닉 확률 분포와 조합을 통해 원래의 음원을 모델링하는 데 사용될 수 있다. 즉, 주파수 및 시계열 성분은 기존 PCA 나 NMF[11]와 같은 행렬 인수분해 방법에서 사용되는 기저 행렬(basis matrices) 및 계수 행렬(coefficient matrices)와 유사한 역할을 가진다. 따라서 분해된 주파수 성분은 분해 음악 신호 전체를 대표하는 스펙트럼 템플릿(spectrum template)으로 간주할 수 있다.

그러나 각 주파수 성분을 그대로 색인화할 경우, 큰 색인 크기로 인해 색인화 효율이 떨어질 수 있다. 음악 등의 오디오 신호에서 주파수 도메인 특성을 추출하기 위해 오디오 신호에 short-time Fourier transform (STFT)를 적용하여 스펙트로그램을 추출할 때, 단위 창문 함수 $\omega(n)$의 길이는 일반적으로 1024 samples 이다. 따라서, 하나의 노래로부터 생성되는 색인은 N개의 512차원 벡터 $P(f|z)$로 표현된다. 한편, 고차원 특성은 많은 수의 은닉 변수 수 및 분해 대상 노래의 검색 효율을 떨어뜨릴 수 있다. 그러므로 효율적인 검색을 위해 추출된 주파수 성분을 축소할 필요가 있다.

본 연구에서는 주파수 성분 축소를 위해 음악 정보 검색 분야에서 널리 쓰이며 그 성능이 입증된 다양한

<표 1> 주파수 성분 축소를 위한 음향 기술자

약어	음향 기술자명 및 설명
	수식 정의
SC	Spectral centroid; 스펙트럼 무게 중심
	$\left(\sum_{k=0}^{N/2} f_k \lvert P(k)\rvert^2\right) / \left(\sum_{k=0}^{N/2} \lvert P(k)\rvert^2\right)$
SS	Spectral spread; 스펙트럼 확산도
	$\sqrt[2]{\left(\sum_{k=0}^{N/2} (f_k - SC)^2 \lvert P(k)\rvert^2\right) / \left(\sum_{k=0}^{N/2} \lvert P(k)\rvert^2\right)}$
SM	Spectral mean; 스펙트럼 에너지 평균
	$\frac{2}{N}\sum_{k=0}^{N/2} P(k)$
SSD	Spectral standard deviation; 에너지 편차
	$\sqrt[2]{\frac{2}{N}\sum_{k=0}^{N/2} (P(k) - \mathrm{SM})^2}$
SF	Spectral flatness; 스펙트럼 평활도(平滑度)
	$\sqrt[2/N]{\prod_{k=0}^{N/2} \lvert P(k)\rvert^2} \;/\; \left(\frac{2}{N}\sum_{k=0}^{N/2} \lvert P(k)\rvert^2\right)$
SSk	Spectral skewness; 스펙트럼 왜도(歪度)
	$\left(\frac{2}{N}\sum_{k=0}^{N/2} (P(k) - \mathrm{SM})^3\right) / \mathrm{SSD}^3$
SK	Spectral kurtosis; 스펙트럼 첨도(添度)
	$\left\{\left(\frac{2}{N}\sum_{k=0}^{N/2} (P(k) - \mathrm{SM})^4\right) / \mathrm{SSD}^4\right\} - 3$
SRF	Spectral roll-off frequency; 롤오프 주파수
	$\arg\min_n \sum_{k=0}^{n} \lvert P(k)\rvert = \alpha \sum_{k=0}^{N/2} \lvert P(k)\rvert$
SB	Spectral brightness; 스펙트럼 밝기
	$\sum_{k=\beta}^{N/2} \lvert P(k)\rvert$

음향 기술자 중 주파수 도메인 특성들을 계산한다. 이러한 접근은 주파수 성분이 주파수 도메인으로 표현되어 있기에 유효하다. 사용하는 음향 기술자의 상세는 표 1 에 설명되어 있다.

이와 같은 음향 기술자를 사용한 특성 추출을 통해 색인의 크기를 줄일 수 있다. 표 1 의 9가지 음향 기술자를 사용하여 하나의 노래로부터 특성을 추출할 경우 N개의 9차원 특성 벡터 $\boldsymbol{x}_m$가 추출된다. 이는 기존의 방법론에 의해 추출된 N개의 1024차원 특성 벡터보다 현저히 적은 규모이다.

3.3. 검색 방법

추출된 음악 요소 특성 벡터의 각 차원별 데이터 분포는 서로 다른 특성 추출 방법을 사용하기에 다를 수 있다. 또한 추가되는 데이터에 따라 데이터베이스 내 표본 집합의 분포는 변화할 수 있다. 따라서 표본 집합 분포 변화에 강인한(scale-invariant) 거리 측정 방법이 필요하다. 이러한 문제를 해결하기 위해, 본 연구에서는 마할라노비스 거리를 사용한다.

마할라노비스 거리(Mahalanobis distance)[12]는 표본 집합과 표본 또는 표본 집합으로 형성된 공간에서 표본 간의 유사도를 구하는 방법으로, 표본집합의 공분산 $\boldsymbol{S}$에 대해 다음과 같이 정의된다:

$$MahalanobisDistance(\boldsymbol{x_1}, \boldsymbol{x_2}) = \sqrt{(\boldsymbol{x_1} - \boldsymbol{x_2})^{\mathrm{T}} \boldsymbol{S}^{-1} (\boldsymbol{x_1} - \boldsymbol{x_2})}$$

한편, 노래로부터 추출된 특성 벡터로 이루어진 표본 집합이 충분히 클 경우, 질의 요소로부터 추출된 특성 $\boldsymbol{x}_q$는 표본 집합의 분포를 따를 수 있다. 따라서 동일한 공분산 $\boldsymbol{S}$를 사용하여 노래 요소로부터 추출된 특성 $\boldsymbol{x}_m$와 특성 $\boldsymbol{x}_q$간 거리를 측정할 수 있다. 이러한 특성 간의 유사도를 이용하여, 노래와 질의로부터 추출된 서로 다른 특성 집합 $\mathbf{X}_m = \{x_1, x_2, \cdots, x_{N_m}\}$ 과 $\mathbf{X}_q = \{x_1, x_2, \cdots, x_{N_q}\}$ 간 유사도를 정의할 수 있다:

$$Similarity(\mathbf{X}_m, \mathbf{X}_q) = \frac{1}{N_m N_q} \Sigma_{x_i \in \mathbf{X}_m} \Sigma_{x_j \in \mathbf{X}_q} MahalanobisDistance(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

(단, N_m 및 N_q는 노래 및 질의의 특성 벡터 개수)

이렇게 정의된 유사도는 노래 및 질의로부터 추출된 특성 간의 유사도로 간주하여 사용할 수 있다.

4. 실험 결과

본 장에서는 기존 특성 추출 방법 및 제안된 방법을 효율성 및 효과성 측면에서 비교하기 위한 다양한 실험을 수행하고 그 결과를 분석한다.

4.1. 실험 환경 및 데이터

전체 실험은 Intel Q9550 2.83GHz Quad Core, 8GB RAM, 4 대의 1TB RAID-5 HDD array 로 구성된 컴퓨팅 환경에서 이루어졌다. 모든 구현 및 실험은 Windows 7 OS 의 MATLAB 2010a 에서 이루어졌다. 구현된 PLCA 및 음향 기술자 추출 모듈을 통해 비교 대상곡 및 질의의 특성을 추출하였다. 이후, 분석된 질의를 전달하여 검색 결과를 출력하였다.

실험용 노래 데이터로는 최근 17 년간 발매된 유명 팝 음악으로 구성된 그래미상 후보(Grammy nominees) 컴필레이션 앨범 1995 년부터 2011 년까지 노래 295 곡을 사용하였다. 노래 전체의 재생시간은 총 19 시간 56 분 55 초이며, 평균 재생시간은 4 분 3.44 초이다.한편, 실험용 질의 데이터로는 남성 및 여성 목소리 각 2 건으로, 총 재생시간은 17 분 58 초이다.

데이터베이스 내 노래 및 질의 분해를 위한 은닉 성분의 수는 전체 시스템에 영향을 미치는 중요한 요소이다. 본 연구에서는 질의보다 노래가 상대적으로 더 많은 정보를 포함하고 있을 것이라는 가정 하에, 질의의 은닉 성분 수는 20, 노래의 은닉 성분 수는 100 으로 설정하였다. 또한, 최대 색인 생성 시간을 제한하기 위해, 확률적 은닉 성분 분석의 EM 알고리즘 반복 횟수를 100 회로 제한하였다.

4.2. 목소리 성별에 의한 내용 기반 질의 결과

성능 평가를 위해 목소리 신호 중 연속된 30 초 구간을 무작위로 추출하는 방법으로 남성 및 여성 목소리 질의를 각각 30 개 생성하였다.

그림 2 는 목소리 성별에 의한 평균 정확도-재현도 (average precision-recall) 그래프이다. 양쪽 모두 아무런 방법을 적용하지 않은 결과(baseline)보다 좋은 성능을 보이고 있으며, 음향 기술자를 추출하여 검색하는 쪽이 좀더 좋은 정확도-재현도 성능을 보이고 있음을 알 수 있다.

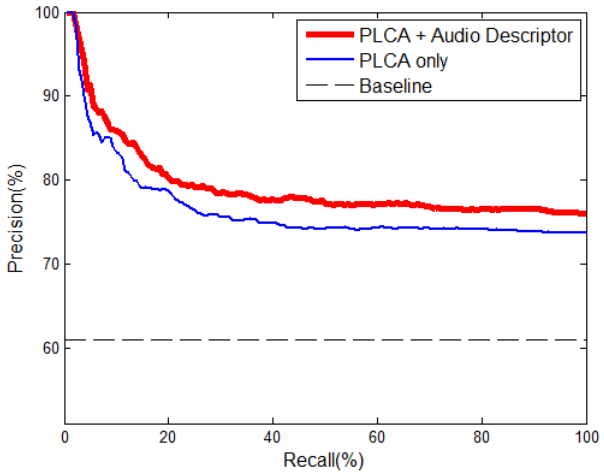
한편, 음향 기술자의 사용 유무에 따른 검색시간도 중요한 이슈이다. 검색에 소요되는 시간은 3.3 장의 마할라노비스 거리 및 특성 집합 간 비교에 따라 달라진다. 기존 음향 기술자를 적용하지 않고 주파수 성분만을 이용한 경우, 질의당 평균 22.74±0.01 초 소요되었다. 반면, 음향 기술자를 이용한 검색의 경우 평균 4.33±0.01 초 소요되었다.

이러한 정확도-재현도 및 검색 시간 결과를 통해, 제안된 방법이 효율 및 효과의 양 측면에서 검색 성능이 향상되었음을 알 수 있었다.

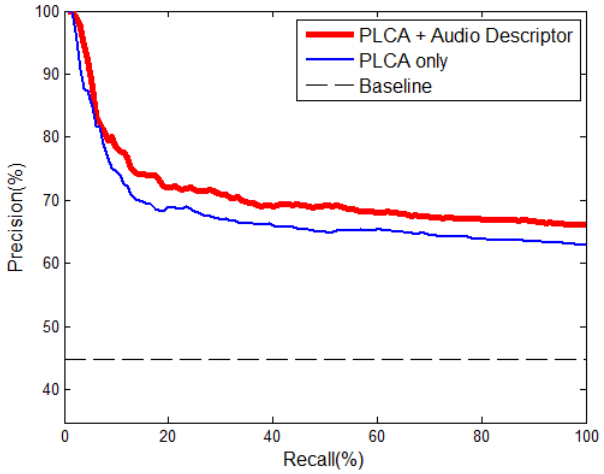
5. 결론

지금까지 내용 기반 음악 검색의 일반화를 위한 내용 기반 음악 요소 검색(CBMER) 방법을 소개하였다. 제안 방법에서는 확률적 은닉 성분 분석(PLCA)을 사용하여 음원을 분해하고, 각 분해 요소로부터 오디오 특성을 추출하였다. 다양한 질의 방식 및 검색 카테고리로 내용 기반 음악 요소 검색이 가능함을 보이기 위해, 남성 및 여성 목소리로부터 질의를 생성하여 음악을 검색하는 실험을 수행하였다. 실험 결과, 제안 방법은 기존의 방법보다 속도 및 정확도-재현도 측면에서 좋은 성능을 보임을 알 수 있었다.

향후 좀더 빠른 내용 기반 음악 요소 분해 및 정확도 높은 음악 요소 검색 성능을 달성하기 위하여 다양한 검색 알고리즘을 적용할 계획이다.



(a)



(b)

(그림 2) (a) 남성, (b) 여성 목소리에 의한 질의 결과

6. 참고문헌

[1] 김무영, 이석필, "MIREX 기술 동향," 전자공학회지, 제 37 권 제 1 호, pp.88-102, 2010 년 1 월.

[2] Paris Smaragdis, Bhiksha Raj, and Madhusudana, Shashanka, "A probabilistic latent variable model for acoustic modeling," *Advances in models for acoustic processing workshop, Neural Information Processing Systems (NIPS)*, Dec. 2006.

[3] 한병준, 노승민, 황인준, "A threshold adaptation based voice query transcription scheme for music retrieval," 한국전기전자학회 전기학회논문지, vol.59, no.2, pp.445-451, 2010 년 2 월.

[4] Jyh-Shing Roger Jang, Hong-Ru Lee, Chia-Hui Yeh, "Query by tapping: A new paradigm for content-based ,music retrieval from acoustic input," *IEEE Pacific Rim Conference on Multimedia 2001*, pp.590-597, Oct. 2001.

[5] Byeong-jun Han, Seungmin Rho, Sanghoon Jun, and Eenjun Hwang, "Music emotion classification and content-based music recommendation," *Multimedia Tools and Applications (MTAP)*, vol.47, no.3, pp.433-460, Aug. 2010.

[6] Seungmin Rho, Byeong-jun Han, Eenjun Hwang, "SVR-based music mood classification and context-based music recommendation," *ACM Multimedia 2009*, pp.713-716, Oct. 2009.

[7] Kyogu Lee and Malcolm Slaney, "Acoustic chord transcription and key extraction from audio using key-dependent HMMs trained on synthesized audio," *IEEE Transactions on Audio, Speech, and Language Processing(ASLP)*, vol.16, no.2, pp.291-301, Feb. 2008.

[8] 한병준, 김연주, 이장우, 김민제, 이교구, "확률적 은닉 성분에 기반한 드럼 onset 검출 방법," 제 34 회 한국정보처리학회 추계학술발표대회, pp.762-765, 2010 년 11 월.

[9] 한병준, 이장우, 이교구, "확률적 은닉 성분 군집화에 기반한 타악기 음원 분리," 2010 년 대한전자공학회 추계학술대회, pp.383-384, 2010 년 11 월.

[10] Namgook Cho and C.-C. Jay Kuo, "Sparse music representation with source-specific dictionaries and its application to signal separation," *IEEE Trans. on ASLP*, vol.19, no.2, pp.337-348, Feb. 2011.

[11] D. D. Lee and H. S. Seung, "Algorithms for nonnegative matrix factorization," *NIPS*, pp.556-562, 2001.

[12] P.C. Mahalanobis, "On the generalised distance in statistics," *in Proceedings of the National Inistitute of Sciences of India*, vol.2, no.1, pp.49-55, 1936.