다중 단계 NMF-EM 알고리즘 기반의 오디오 소스 분리 방법에 대한 연구

조충상, 김제우
전자부품연구원
ideafisher@keti.re.kr, jwkim@keti.re.kr

A sturdy on the blind audio source separation based on multi-step NMF-EM algorithm

Choongsang Cho and Jewoo Kim
Korea Electronics Technology Institute

요 약

본 논문에서는 오디오 신호의 특성 표현에 유용한 nonnegative matrix factorization(NMF)에 대해 설명하였으며, expectation maximization (EM)을 이용한 NMF 파라미터 추출 및 EM-NMF 기반한 오디오 소스 분리 기술에 대해서 설명했다. 또한, 다중 단계 NMF-EM 구조의 객체 분리를 통해서 객체 분리 성능을 향상시키기 위한 알고리즘을 제안하며, 제안된 알고리즘은 K-pop 음원과 SDR(source distortion ratio)를 통해서 객체 분리 성능을 평가한다. 성능 평가 결과 제안된 알고리즘은 다중 단계를 통해 약 3dB 의 보컬 분리 성능이 향상되며, 상업적 음원 제작에서 사용되는 가상 오디오 효과가 많이 적용된 음원에서 약 5dB 의 분리 성능을 향상시켰다. 그러므로 제안된 방식은 오디오 객체 분리에 유용한 방법이 될 것으로 생각된다.

1. 서론

오디오 소스 분리에 대한 연구는 오랫동안 다양한 방법으로 폭넓게 이루어지고 있며, 최근에는 오디오 음원 분리 기술을 기반으로 기존에 제작된 미디어 콘텐츠의 배경음 수정, 오디오 변환 및 3D 오디오 효과 등 다양한 분야에 응용되고 있다. 오디오 소스 분리에 대한 연구는 다양한 접근 방법들을 통해 연구 되고 있다. 첫 번째 방법은 사전 정보 없이 수행되는 오디오 분리 기술로서, 오디오 소스들이 서로 독립적인 성질을 갖는다는 가정을 기반으로 신호의 독립성이 최대가 되도록 하는 independent component analysis(ICA)가 있으며[1], 두 번째 방법은 singular value decomposition(SVD) 혹은 주파수 도메인에서 관찰되는 채널간의 독립적인 정보를 이용하여 오디오 소스들을 믹싱한 시스템에 추정하여, 추정된 시스템의 역 시스템을 이용하여 오디오 객체 음원을 분리하는 방식이 있다[1]. 세 번째로는 사전 정보를 활용하는 방식으로 사전에 취득한 오디오 객체별 특성정보를 바탕으로 최적화된 파라미터와 믹싱시스템 파라미터를 추정하는 방식이 있다[2,3]. 특히 최근에 오디오 객체의 특성 정보를 추정하는 방식으로 멀티 매트릭스 구조를 이용한 nonnegative matrix factorization(NMF)이 널리 사용되고 있으며, 이를 바탕으로 NMF-(expectation maximization) EM 을 탑재한 오디오 소스분리 방식이 연구되고 있다[2,3]. 본 논문에서는 NMF-EM 기반의 오디오 소스 분리 기술에 대해서 설명하고, NMF-EM 의 다중 단계 구조를 이용하여 객체를 분리하는 방식을 제안한다.

2. NMF-EM 기반의 오디오 객체 분리

일반적으로 오디오 소스 분리 기술은 믹싱된 오디오 신호에서 믹싱에 사용된 오디오 소스들을 분리하는 것이다. 이러한 구조를 수식적으로 간단히 표현하면 수식 1 과 같이 소스 신호 S 가 믹싱 시스템 A 에 적용되고, 노이즈 n 가 추가되어 얻어진 믹싱 신호 $\mathbf{x}$ 로 표현된다.

$$\mathbf{x}=AS+n \quad (1)$$

오디오 믹싱시스템에서 믹싱된 소스의 개수 J 가 관찰 가능한 채널 I 수 보다 클 경우, 미지수 보다 구성할 수 있는 수식의 수가 적으므로, 수식 1 은 ill pose inverse problem 으로 믹싱 시스템 A 를 추정하기 위해서는 통계적인 방법, 혹은 사전 정의나 정보가 필요하다.

오디오 소스 분리 기술은 믹싱된 신호 $\mathbf{x}$ 만을 관찰할 수 있는 상황에서 소스신호 S 를 찾아내는 방법으로, 일반적인 사전정보 활용한 오디오 소스 분리 기술은 오디오 소스들에 대한 특성 파라미터 추출 단계, 특성 파라미터 및 믹싱 파라미터 예측단계와 오디오 소스 분리 단계로 구성된다. 오디오 소스 분리 기술에서는 오디오 신호의 특성을 NMF 를 이용하여 많이 사용하고 있다[2]. NMF 기술은 오디오 신호 $\mathbf{x}$ 의 스펙트럼 데이터를 H 와 W 두 개의 매트릭스 곱으로 표현되도록 하는 방식으로, short-term fourier transform(STFT)으로 주파수 도메인으로 변환한 후 오디오

신호이 최적의 매트릭스 곱으로 표현되도록 수식 2 와 같이 Gaussian 분포 가정에서 log likelihood 가 최대가 되도록 하는 두 개의 매트릭스를 구하면 된다.

$$C_{ML}(W,H) \overset{def}{=} -\log\left(p(X \mid WH)\right) \quad (2)$$

여기서 $X = STET(\mathrm{x})$ 는 X 의 스펙트럼 신호이다.

NMF-EM 기반의 오디오 분리 기술은 각 소스에 대한 특성 파라미터 H 와 W 을 수식 2 를 기반으로 획득된다.

객체 특성파라미터 기반에서 오디오 소스를 분리하기 위해서는 획득된 특성파라미터를 바탕으로 최적의 믹싱 시스템을 도출하는 모델 예측 작업을 수행해야 한다. 이를 위해서 노이즈 신호에 대한 가정을 추가하게 된다. 먼저 노이즈 신호는 $n = X - AS$ 로 표현 가능하며, 노이즈 신호를 평균값이 0 인 Gaussian 신호로 가정하여 likelihood 를 최대화 하는 방식을 통해서 최적의 객체별 특성 파라미터와 믹싱 시스템을 찾는다. 그러므로 평균이 0 인 Gaussian 분표를 갖는 신호는 수식 3 과 같이 표현 가능하다.

$$N(0,\Sigma_b) = \frac{1}{\sqrt{2\pi}^{mn}\Sigma_b^{\,m/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{m}(x_i)^T\Sigma_b^{-1}(x_i)\right] \quad (3)$$

여기서 Σ_b 는 노이즈의 변화량을 나타낸다.

노이즈에 대한 수식 2 를 수식 3 에 적용하면 수식 4 와 같은 Gaussian 분포로 표현 가능하다.

$$N(0,\Sigma_b) = \frac{1}{\sqrt{2\pi}^{mn}\Sigma_b^{\,m/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{m}(x-AS)^T\Sigma_b^{-1}(x-AS)\right] \quad (4)$$

노이즈에 대한 확률 분포 모델이 수식 4 와 같이 정의 되었을 경우, 확률 모델을 구성하는 파라미터가 최적의 파라미터를 갖도록 하기 위하여 수식 5 와 같이 기대값을 최대화(Expectation Maximization)하는 방식을 통해 획득할 수 있다[3].

$$\begin{aligned} Q(\theta \mid \theta') &= E\left[-\log p(X, C \mid \theta) \mid X, \theta'\right] \\ &= E\left[-\log p(X \mid C, \theta) - \log p(C \mid \theta) \mid X, \theta'\right] \end{aligned} \quad (5)$$

여기서 $S = \sum_{k \in K} c_{k,fn}$ 이고, θ 는 EM 알고리즘의 파라미터이고, $W, H, A \in \theta$ 이다.

획득된 모델 파라미터는 Winner filter 의 계수로 사용되며, 이를 통해서 오디오 소스가 분리되게 된다[3].

3. 다중 NMF-EM 기반의 오디오 객체 분리

본 연구에서는 NMF-EM 기반의 객체 분리 알고리즘의 성능을 개선하기 위하여, 그림 1 과 같이 사전에 각 각체별 연산된 NMF 데이터를 NMF-EM 알고리즘 기반 오디오 객체 분리에 적용하여, 믹싱된 오디오 데이터를 1 차 오디오 객체 분리를 수행한다. 1 차 객체 분리를 통해서 획득된 데이터 $O_1, \ldots, O_N$ 에는 각 객체의 정보가 주요하게 포함되어 있으며, 다른 객체의 정보도 포함되어 있다. 그래서 본 연구에서 개발된 방식은 1 차 분리된 결과를 다시 분리하여 O_i 에 포함되어 있는 다른 객체 해당하는 성분을 분리하기 위해, 2 차 오디오 객체 분리가 수행되며 이를 통해 분리된 객체 데이터 S_i 가 획득된다. 만약 1 차 분리에서 N 개의 객체로 분리되면, 2 차 분리에서는 N 개의 NMF-EM 객체 분리 모듈이 추가적으로 필요하다. 이러한 구조는 전체 객체분리의 복잡도를 상당히 증가시킨다. 그러므로 본 연구에서는 믹싱된 음원에서 청취자가 가장 집중하게 되는 보컬 성분에 대해서 2 차 NMF-EM 기반의 오디오 객체 분리를 수행하며, 초기 객체들에 대한 NMF 파라미터 값은 1 차 분리 모듈에서 사용된 동일한 값을 사용한다.

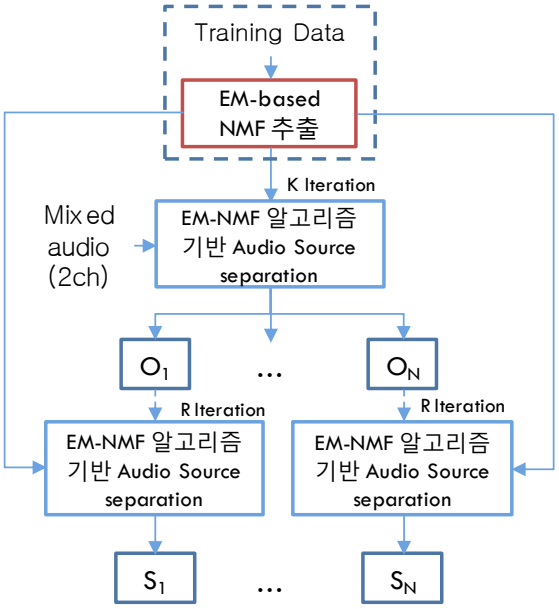


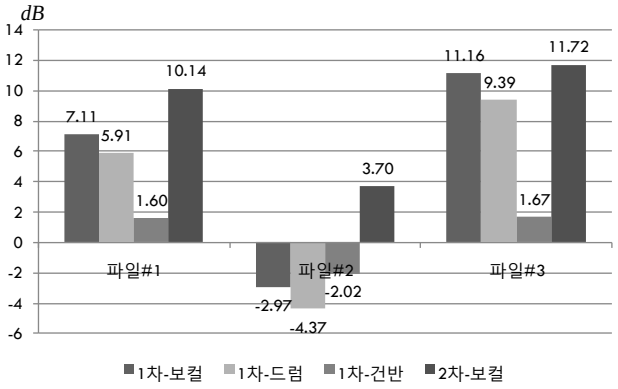
그림 1. 다중 단계를 이용한 객체분리 구조도

3. 실험

본 논문에서는 NMF-EM 기반의 다중 단계 분리 알고리즘을 평가하기 위하여, 객체오디오 서비스를 위해 출시되었던 Kpop 음원을 이용하여 성능 평가를 수행한다. 성능 평가에 사용된 음원은 44.1kHz, 16bits, 스테레오로 녹음된 음원으로 각 객체에 대한 독립적인 음원이 제공되며, 믹싱된 음원을 생성하기 위하여 K-pop 음원의 보컬, 드럼, 건반 악기를 임의의 비율로 믹싱하여 믹싱된 음원을 생성하였다. NMF-EM 기반의 알고리즘은 초기 NMF 파라미터 값을 요구한다. 본 실험에서 NMF 파라미터를 획득하는 방법에 대한 영향을 최소화하여 오디오 분리 방법에 대한 평가의 정밀도를 높이기 위해서, NMF 파라미터는 믹싱된 음원의 객체별 데이터를 이용해서 EM 기반의 NMF 추출 방법에 적용하여 각 객체별 NMF 파라미터를 획득한다[2]. 또한 NMF 파라미터 추출 단계에서 NMF 차원은 각각 {6, 4, 4}으로 설정되었으며, EM 최적화 단계는 각각 1000 의 반복적 수행이 이루어졌다. 성능 비교를 위해서 NMF-EM 알고리즘의 기반의 객체 분리 알고리즘과 비교되었으며, NMF 초기 파라미터는 제안된 방식과 비교된 방식이 모두 동일한 파라미터를 사용하였다. 실험에서 제안된 방식의 첫 번째 단계에서 EM 기반의 최적화 동작이 200 번의 반복적 수행이 이루어지고, 두 번째 단계에서는 보컬에 대해서만 2 차 분리가 이루어지며 30 번의 EM 최적화 동작이 이루어진다.

표 1 은 분리된 객체 음원과 원본 객체 음원을 이용하여 획득된 SDR(Source Distortion ratio) 결과를 나타낸다. 실험 파일 중 파일#2 는 상업적 음원제작에서 많이 사용되는 가상 오디오 효과들이 많이 사용된 음원으로, 다른 음원에 비해서 낮은 객체분리 성능을 보인다. 그리고, 제안된 2 차 분리를 사용하면서 보컬에서 1 차 분리 결과보다 약 3dB 의 상승의 보이며, 가상효과가 많이 사용된 파일#2 에서 약 5dB 정도의 객체분리 성능이 향상되는 것을 확인하였다.

표 1. 다중 단계를 이용한 객체 분리 성능



4. 결론

본 논문에서는 NMF 파라미터를 통하여 오디오 신호를 분석하는 방법과 NMF 파라미터와 EM 을 사용한 오디오 소스 분리 방법을 분석하였고, 다중 단계를 이용한 오디오 객체 분리 방법을 제안하였다. 성능 평가를 위한 상업적 목적으로 제작된 Kpop 음원을 사용하였으며, 성능 평가를 위해 SDR 를 측정하였다. 측정 결과 제안된 방식은 보컬 2 차 분리단계를 통해서 약 3dB 정도 성능 향상을 보였으며, 특히 상업 음원 제작에서 많이 사용되는 가상 오디오 효과들에 의해 낮은 분리 성능을 보이는 파일에서도 약 5dB 의 성능향상을 보였다.

그러므로 본 논문에서 제안된 NMF-EM 기반의 다중 단계 오디오 객체분리 구조는, 오디오 분리 알고리즘의 유용한 방법으로 사용될 수 있을 것으로 생각된다.

본 연구는 미래창조과학부 및 한국산업기술평가관리원의 산업융합원천기술개발사업의 일환으로 수행하였음. [10044569, 2D 스테레오 콘텐츠를 3D 입체 음향 콘텐츠로 변환하기 위한 음원 객체 분리/위치 추정 및 3D 렌더링 소프트웨어 기술 개발'.

참 고 문 헌

[1] Emmanuel Vincent, "Complex nonconvex lp norm minimization for underdetermined source separation," In Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA), pp. 430-437, 2007.

[2] C´edric F´evotte , Nancy Bertin, and Jean-Louis Durrieu, "Nonnegative Matrix Factorization with the Itakura-aito Divergence: With Application to Music Analysis," Neural Computation, vol. 21, pp. 793-830, 2009.

[3] Alexey Ozerov and Cédric Févotte, " Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation," IEEE Trans. Audio Speech and Language Proc., vol. 18, no. 3, pp. 550-563, Mar. 2010.

[4] P. Bofill, "Identifying Single Source Data for Mixing Matrix Estimation in Instantaneous Blind Source Separation," In proc. of ICANN 2008, pp. 759-767, 2008.