EADRess 와 NMF 를 결합한 음원분리 성능 분석

정영호, 장대영, 이태진
한국전자통신연구원
yhcheong@etri.re.kr, dyjang@etri.re.kr, tjlee@etri.re.kr

Performance Analysis of Sound Source Separation Combining EADRess and NMF

Youngho Jeong, Daeyoung Jang, Taejin Lee
Electronics & Telecommunications Research Institute

요 약

본 논문에서는 스테레오 채널 신호 간 강도비를 이용하여 음원을 분리하는 EADRess 알고리즘과 부분기반 표현을 특징으로 한 비음수 행렬 인수분해를 통해 음원을 분리하는 NMF 가 결합된 새로운 음원분리 알고리즘을 제안한다. 입력 오디오 신호로부터 frequency-azimuth 평면 구성을 통해 식별된 방위각에 상응하는 신호 강도비로 표현되는 확률밀도함수를 이용하여 1 단계 음원분리를 수행하고, 얻어진 개별 분리음원을 대상으로 supervised NMF 및 Wiener 필터 기반 마스킹 함수를 적용함으로써 잔류 혼합성분을 제거하는 2 단계 음원분리를 수행한다. 제안된 EADRess/NMF 결합 음원분리 알고리즘의 성능을 검증하기 위하여 SASSEC 에서 제공하는 테스트 음원을 이용하여 측정한 결과, 개별 음원분리 알고리즘에 비해 SIR 이 각각 1.41dB, 10.43dB 향상된 결과를 얻었다.

1. 서론

인간의 좌/우 귀에 입력되는 오디오 신호 간 강도 차(IID: Inter-aural Intensity Difference)를 기반으로 음원의 위치를 인지하는 인간의 청각 특성을 이용하는 대표적인 음원분리 기술로 ADRess(Azimuth Discrimination and Resynthesis) 알고리즘이 있다[1]. 대부분의 상용 오디오 콘텐츠가 개별 음원들을 IID 기반으로 패닝하고 이를 믹싱하여 제작되는 점을 감안한다면, ADRess 알고리즘은 스테레오 오디오 신호 기반 음원분리에 그 활용 가능성이 매우 크다고 할 수 있다.

또한 부분기반 표현(parts-based representation)을 바탕으로 객체를 인식하는 뇌의 인지 특성을 이용하는 NMF(Non-negative Matrix Factorization) 알고리즘은 음원분리, 음성향상, 배경음 추출 등의 다양한 신호처리 분야에 적용되고 있다[2][3]. 본 논문에서는 이와 같은 인간의 청각 및 뇌의 인지 특성을 기반으로 한 EADRess(Enhanced ADRess) 알고리즘과 NMF 알고리즘을 결합한 새로운 음원분리 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2 절에서는 ADRess 알고리즘의 성능을 개선한 EADRess 알고리즘의 음원 방위각 식별 및 분리음원 합성 방법에 대해 살펴보고, 3 절에서는 NMF 알고리즘의 비음수 행렬 인수분해를 위한 업데이트 규칙 유도과정을 설명한다. 4 절에서는 본 논문에서 제안한 EADRess/NMF 결합 알고리즘의 처리과정을 기술한다. 5 절에서는 SASSEC(Stereo Audio Source Separation Evaluation Campaign) 테스트 음원 및 SIR 평가 지표를 이용한 성능분석 결과를 보여주고, 마지막으로 6 절에서는 본 논문에 대한 결론을 맺는다.

2. EADRess 알고리즘

기존 ADRess 알고리즘의 성능을 개선한 EADRess 알고리즘은 음원 방위각 식별 및 분리음원 합성의 두 단계로 구성되며, 각 처리 단계별 상세 내용은 다음과 같다.

먼저, 음원 방위각 식별을 위해 매 신호 분석 프레임에 대해 STFT 를 취하고, 식(1)을 이용하여 $(N+1)\mathrm{x}(\beta+1)$ 배열의 frequency-azimuth 평면을 구성한다. N 과 β 는 각각 주파수 해상도와 방위각 해상도를 의미한다.

$$A_Z(k,m,i)=\begin{cases}|X_r(k,m)-g(i)X_l(k,m)| & if\ i\le\beta/2\\ |X_l(k,m)-g(i)X_r(k,m)| & if\ i>\beta/2\end{cases} \quad (1)$$

여기서, k는 $0\le k\le N$를 만족하고, $X_l(k,m)$과 $X_r(k,m)$는 각각 좌측과 우측 채널의 m번째 프레임에서의 k번째 주파수 성분을 나타낸다.

Sinusoidal energy-preserving panning law 를 기반으로 좌/우 채널간 신호 강도비 $g(i)$는 식(2)와 같이 정의되며, 0 과 1 사이의 값을 가진다.

$$g(i)=\begin{cases}e^{log\left(tan\left(i\frac{\pi}{360}\right)\right)} & if\ i\le\beta/2\\ e^{log\left(tan\left((180-i)\frac{\pi}{360}\right)\right)} & if\ i>\beta/2\end{cases} \quad (2)$$

여기서 i는 $0\le i\le\beta$, i와 β는 정수이다. β 값이 커질수록 방위각 해상도를 높일 수는 있으나, 계산량은 증가하게 되므로

이를 감안하여 설정한다. 이때 방위각은 $180\left({}^{i}/_{\beta}\right)$로 표현된다.

음원이 좌측 채널에서 우세한 경우$(i \le \beta/2)$, 그리고 우측 채널에서 우세한 경우$(i > \beta/2)$에서 A_Z이 최소가 되는 $g(i)$를 찾을 수 있다. 해당 위치에서의 음원 에너지를 추정하기 위해 식(2)를 식(3)으로 재정의하고, 이를 토대로 frequency-azimuth 평면을 구성한다.

$$A_{\bar{Z}}(k,m,i) = \begin{cases} A_Z(k,m)_{max} - A_Z(k,m)_{min} & if\ A_Z(k,m,i) = A_Z(k,m)_{min} \\ 0 & otherwise \end{cases} \quad (3)$$

여기서, max 와 min 은 k 번째 주파수 성분에 대한 azimuth 축 선상의 $A_Z(k,m,i)$ 최대값과 최소값을 의미한다.

구성된 평면에서 방위각을 기준으로 누적된 $A_{\bar{Z}}(k,m,i)$ 값을 이용하여, 분리하고자 하는 음원의 개수만큼 peak 값들을 찾음으로써 음원의 정확한 방위각 d_j를 구할 수 있다.

식(2)의 신호 강도비 $g(i)$는 방위각 90°를 중심으로 좌우 대칭값을 갖는 관계로 좌/우측 방위각에 대한 패닝 모호성이 발생하며, 이를 해결하기 위해 식(4)와 같이 $\bar{g}(i)$를 정의한다.

$$\bar{g}(i) = \begin{cases} (1 - g(i)) \cdot (-1) & if\ i \le \beta/2 \\ 1 - g(i) & if\ i > \beta/2 \end{cases} \quad (4)$$

분리음원 합성을 위해 식(6)과 같이 식별된 방위각에 상응하는 신호 강도비로 표현되는 확률밀도함수인 가우시안 윈도우 함수 $G_j(k,m)$를 정의한다.

$$U(k) = \underset{0 \le i \le \beta}{\arg\min} A_Z(k,m,i) \quad (5)$$

$$G_j(k,m) = \frac{1}{\sqrt{2\pi\gamma}} e^{-\left(\bar{g}(U(k)) - \bar{g}(d_j)\right)^2/(2\gamma)} \quad (6)$$

여기서, γ 는 윈도우 폭을 제어하며, 분리음원의 왜곡 성분이 커지지 않도록 적절히 설정되어야 한다. $U(k)$ 는 식(5)에서와 같이 매 신호 분석 프레임별로 k 번째 주파수 성분에서 $A_Z(k,m)_{min}$ 을 갖는 인덱스 i 값이다.

식(7)에서와 같이 앞서 정의한 가우시안 윈도우 함수를 좌/우 신호 중 하나의 주 혼합 신호에 취함으로써 분리음원에 대한 주파수 영역 신호를 구한다.

$$Y_j(k,m) = \begin{cases} G_j(k,m) \cdot X_l(k,m) & if\ d_j \le \beta/2 \\ G_j(k,m) \cdot X_r(k,m) & if\ d_j > \beta/2 \end{cases} \quad (7)$$

3. NMF 알고리즘

NMF 는 다음과 같이 행렬의 원소가 비음수인 특징을 갖는 인수분해를 의미한다.

$$\boldsymbol{V} \approx \boldsymbol{WH} \quad (8)$$

여기서, $\boldsymbol{V}$는 분석하고자 하는 데이터 행렬$(n \times m)$을, $\boldsymbol{W}$는 r개의 기저 벡터로 구성된 기저 행렬$(n \times r)$을, $\boldsymbol{H}$는 기저 벡터의 선형 조합 계수를 포함하는 부호화 행렬$(r \times m)$이다. 이때 모든 행렬 $\boldsymbol{V}$, $\boldsymbol{W}$, $\boldsymbol{H}$ 원소는 비음수이며, 일반적으로 인수분해 차수 r은 $(n+m)r < nm$ 조건하에 선정한다.

관측 데이터 $V_{i\mu}$ 가 평균값이 $(\boldsymbol{WH})_{i\mu}$ 인 Poisson distribution 에 의해 발생하는 모델을 가정하면, $\boldsymbol{WH}$ 에 의해 $\boldsymbol{V}$가 발생할 Poisson likelihood 는 i와 μ에 대해 $log\ P_{i\mu}(\boldsymbol{V}|\boldsymbol{WH})$ 값들을 합함으로써 계산된다. $\boldsymbol{W}$, $\boldsymbol{H}$ 를 최적화 시키기 위한 규칙을 찾기 위해, 상수항을 제거하고 식(9)와 같이 Poisson likelihood 를 최대화시키는 목적함수 D를 정의할 수 있으며, 이는 부호가 반대인 간략화된 Kullback-Leibler divergence 에 해당된다.

$$D(\boldsymbol{V} \parallel \boldsymbol{WH}) = \sum_{i=1}^{n}\sum_{\mu=1}^{m}\left(V_{i\mu} \ln(\boldsymbol{WH})_{i\mu} - (\boldsymbol{WH})_{i\mu}\right) \quad (9)$$

이때 $(WH)_{i\mu}$는

$$V_{i\mu} \approx (WH)_{i\mu} = \sum_{a=1}^{r} W_{ia}H_{a\mu} \quad (10)$$

목적함수 D를 local maximum 으로 수렴시키는 업데이트 규칙을 찾기 위해 목적함수 D를 $\boldsymbol{H}$에 대해 편미분하고,

$$\frac{\partial}{\partial H_{a\mu}} D(V \parallel WH) = \sum_{i=1}^{n} \frac{V_{i\mu}W_{ia}}{(WH)_{i\mu}} - \sum_{i=1}^{n} W_{ia} \quad (11)$$

최적화를 위해 gradient ascent 방식을 적용하고,

$$H_{a\mu} \leftarrow H_{a\mu} + \eta_{a\mu}\left[\sum_{i=1}^{n} \frac{V_{i\mu}W_{ia}}{(WH)_{i\mu}} - \sum_{i=1}^{n} W_{ia}\right] \quad (12)$$

수렴속도를 조절하는 step size 값인 $\eta_{a\mu}$를 식(13)과 같이 설정하면,

$$\eta_{a\mu} = \frac{H_{a\mu}}{\sum_i W_{ia}} \quad (13)$$

$H_{a\mu}$에 대한 multiplicative update rule 인 식(14)를 구할 수 있다.

$$H_{a\mu} \leftarrow H_{a\mu} \frac{\sum_i W_{ia}V_{i\mu}/(WH)_{i\mu}}{\sum_i W_{ia}} \quad (14)$$

마찬가지로 목적함수 D를 $\boldsymbol{W}$에 대해 편미분하여 정리하면, 식(15)와 같은 W_{ia}에 대한 multiplicative update rule 을 구할 수 있다.

$$W_{ia} \leftarrow W_{ia} \frac{\sum_\mu H_{a\mu}V_{i\mu}/(WH)_{i\mu}}{\sum_\mu H_{a\mu}} \quad (15)$$

앞서 유도한 업데이트 규칙을 이용한 NMF 처리 과정은

다음과 같다.

1) 랜덤 양수값으로 $\boldsymbol{W}$, $\boldsymbol{H}$를 초기화 한다.

2) $\boldsymbol{W}$, $\boldsymbol{H}$를 업데이트 하고, 기저 벡터를 정규화한다.

$$W_{ia} \leftarrow \frac{W_{ia}}{\sum_i W_{ia}}$$

3) 수렴할 때까지 1, 2 단계를 반복한다.

4. EADRess/NMF 결합 알고리즘

EADRess 와 NMF 알고리즘을 결합한 스테레오 오디오 신호 기반 음원분리 처리 절차는 그림 1 과 같다.

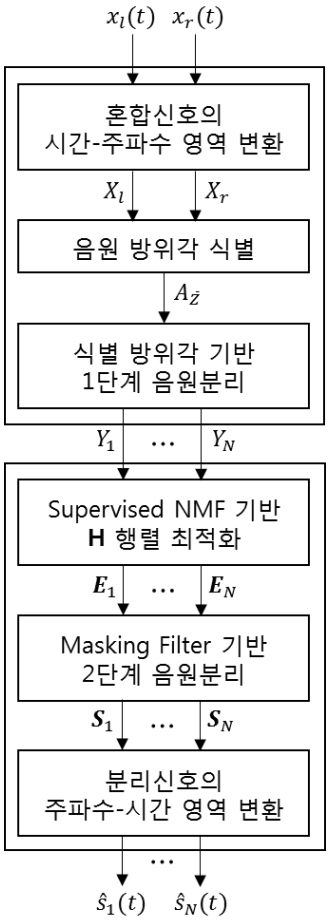


그림 1. EADRess/NMF 결합 알고리즘 처리 절차

그림에서 보는 바와 같이, 2 절에서 설명한 EADRess 알고리즘의 1 단계 음원분리 결과인 Y_j 에 대해, 개별적으로 supervised NMF 를 적용하여 $\boldsymbol{H}_j$ 행렬을 최적화한다. 이때 적용된 supervised NMF 는 기저벡터의 순열 모호성(permutation ambiguity)으로 인한 그룹화 어려움을 해결하기 위해, 분리하고자 하는 음원 정보를 이용한 사전 학습 과정을 통해 기저행렬 $\boldsymbol{W}_j$를 추출하고, 이를 NMF 에 적용하여 $\boldsymbol{H}_j$ 행렬만을 최적화 방식이다.

음원분리를 위해 아래와 같은 Wiener 필터 기반의 masking 함수 $\boldsymbol{M}_j$를 정의한다.

$$\boldsymbol{M}_j = \frac{\boldsymbol{E}_j}{\boldsymbol{WH}} \tag{16}$$

여기서, $\boldsymbol{E}_j$ 는 supervised NMF 에 의해 추정된 j 번째 음원에 대한 스펙트로그램 성분이며, 나누기는 element-wise 연산을 나타낸다.

마스킹 함수가 적용된 주파수 영역에서의 j 번째 최종 분리음원은 다음과 같다.

$$\boldsymbol{S}_j = \boldsymbol{Y}_j \odot \boldsymbol{M}_j \tag{17}$$

여기서, $\odot$는 Hadamard product(element-wise product) 연산을 나타낸다.

ISTFT 를 통해 구한 매 프레임 별 시간 영역에서의 분리음원 신호 $\hat{s}_j(t)$ 는 overlap-add 기법에 의해 재결합됨으로써 주파수-시간 영역 변환 과정을 마무리한다.

5. 실험 결과

본 논문에서 제안한 EADRess/NMF 결합 알고리즘에 대한 성능 분석을 위해 SASSEC 에서 제공하는 테스트 음원 및 객관적 평가 지표를 이용하였다[4].

성능 분석에 사용된 혼합 음원은 서로 다른 방위각(s_1:140°, s_2:100°, s_3:75°, s_4:45°)을 갖는 4 명의 여성 음성을 입력으로 믹싱되었으며, s_1 과 s_2 는 영어, s_3 과 s_4 는 일어로 녹음되었다.

객관적 평가지표로는 SIR(Source to Interference Ratio)를 사용하였으며, 이는 식(12)에서와 같이 추출된 분리음원 $\hat{s}(t)$에 대한 성분 분해를 통해 다음과 같이 정의된다[5].

$$\hat{s}(t) = s_{target}(t) + e_{interf}(t) + e_{noise}(t) + e_{artif}(t) \tag{18}$$

$$SIR = 10\,log_{10}\frac{\|s_{target}\|^2}{\|e_{interf}\|^2} \tag{19}$$

EADRess, NMF(Supervised NMF)와 EADRess/NMF 알고리즘을 대상으로 개별 분리음원에 대한 SIR 성능분석 결과는 표 1 과 같다.

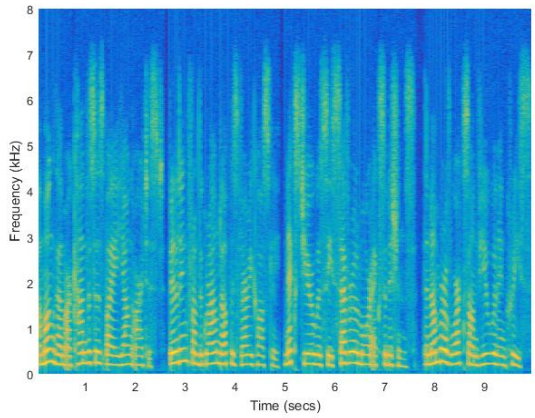
표 1. 개별 분리음원에 대한 SIR 성능분석 결과

	EADRess	NMF	EADRess/NMF
분리음원 $\hat{s}_1$	21.10	9.15	22.17
분리음원 $\hat{s}_2$	19.83	13.50	24.02
분리음원 $\hat{s}_3$	19.09	8.75	20.41
분리음원 $\hat{s}_4$	22.07	14.61	21.12

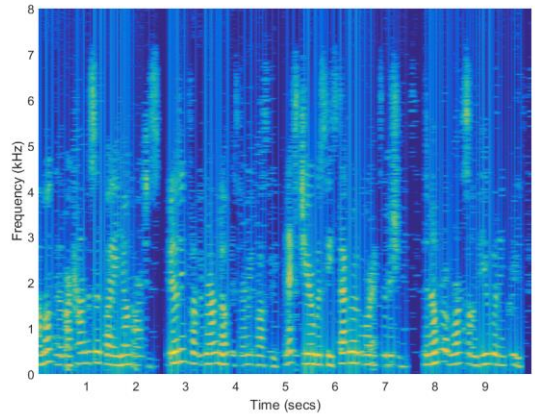
성능분석 결과에서 보는 바와 같이, 본 논문에서 제안한

결합 음원분리 알고리즘은 개별 알고리즘을 단독으로 사용한 경우에 비해 성능이 향상되었다. NMF 알고리즘은 EADRess 알고리즘에 비해 10dB 내외의 성능 열화가 나타났으며, 음성 혼합신호의 음원분리에 사용하는 경우에는 원하는 품질의 분리음원을 얻기 어려울 것으로 보인다. EADRess/NMF 결합 알고리즘의 경우, 대부분의 분리음원에서 EADRess 에 비해 유의미한 성능 개선 효과가 있었다. 그러나 분리음원 $\hat{s}_4$ 에 대해서는 적으나마 성능 열화가 발생하였으며, 이는 EADRess 알고리즘에 의한 1 단계 음원분리가 잘 이루어져 다른 음원 성분이 거의 포함되지 않음으로써 발생한 음질 열화 현상으로 추정된다.

원음원 s_1 과 분리음원 $\hat{s}_1$ 의 스펙트로그램은 그림 2 와 같으며, 그림에서 보는 바와 같이 주요 스펙트럼 성분에 대한 음원분리가 적절히 이뤄졌음을 알 수 있다.



(a) 원음원 s_1의 스펙트로그램



(b) 분리음원 $\hat{s}_1$의 스펙트로그램

그림 2. 원음원 s_1과 분리음원 $\hat{s}_1$ 간 스펙트로그램 비교

표 2. 음원분리 알고리즘 간 SIR 성능분석 결과 비교

	EADRess	NMF	EADRess/NMF
SIR	20.52	11.50	21.93

표 2 는 제안된 EADRess/NMF 결합 알고리즘과 개별 알고리즘 간의 SIR 성능분석 비교 결과를 나타낸다. 제안된 결합 알고리즘이 기존 알고리즘에 비해 개선된 성능을 보였으며, EADRess 에 비해 1.41dB, NMF 에 비해 10.43dB 향상되었다.

6. 결론

본 논문에서는 EADRess 와 NMF 를 결합한 새로운 음원분리 알고리즘을 제안하였다. 이를 위해 입력 오디오 신호로부터 frequency-azimuth 평면을 구성하고, 식별된 방위각에 상응하는 신호 강도비로 표현되는 확률밀도함수를 이용하여 1 단계 음원분리를 수행하였다. 얻어진 개별 분리음원을 대상으로 supervised NMF 및 Wiener 필터 기반 마스킹 함수를 적용함으로써 잔류 혼합성분을 제거하는 2 단계 음원분리를 수행하였다. 제안된 결합 음원분리 알고리즘에 대한 성능 평가는 SASSEC 에서 제공하는 테스트 음원 및 객관적 성능평가 방법을 이용하여 실시하였다. 제안된 결합 알고리즘에 대한 성능분석 결과, 기존 개별 알고리즘에 비해 SIR 이 각각 1.41dB, 10.43dB 향상된 것으로 분석되었다.

감사의 글

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 정보통신 · 방송 연구개발 사업의 일환으로 하였음. [R0126-15-1034, 채널/객체 융합형 하이브리드 오디오 콘텐츠 제작 및 재생기술 개발]

참고문헌

[1] D. Barry et al., “Sound Source Separation: Azimuth Discrimination and Resynthesis,” 7th International Conference on Digital Audio Effects, pp.240-244, Oct. 2004.

[2] D.D. Lee and H.S. Seung, “Learning the Parts of Objects by Non-negative Matrix Factorization,” Nature, vol.401, pp.788-791, 1999.

[3] T. Virtanen, “Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria,” IEEE Trans. on Audio, Speech, and Language Processing, vol.15, no.3, pp.1066-1074, March 2007.

[4] E. Vincent et al., “First Stereo Audio Source Separation Evaluation Campaign: Data, Algorithms and Results,” International Conference on Independent Component Analysis and Signal Separation, pp.552-559, Feb. 2007.

[5] E. Vincent et al., “Performance Measurement in Blind Audio Source Separation,” IEEE Trans. on Audio, Speech, and Language Processing, vol.14, no.4, pp.1462-1469, July 2006.