ADRess 알고리즘 기반 새로운 분리음원 합성 기법

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

New Separated Sound Source Synthesis based on ADRess Algorithm

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

요 약

본 논문에서는 스테레오 오디오 신호를 이용하여 음원을 분리하는 ADRess 알고리즘을 기반으로, 추정된 음원 방위각에 대한 신호 강도비를 이용하여 분리음원을 생성하는 새로운 분리음원 합성 기법을 제안한다. 입력된 스테레오 채널 신호 간 강도 차(IID) 특성을 이용하여 신호 분석 프레임별로 개선된 신호 강도비 함수에 따른 frequency-azimuth 평면을 구성하고, 이를 통해 추정된 방위각에 상응하는 신호 강도비로 표현되는 확률밀도함수를 좌/우 신호 중 하나의 주 입력 신호에 취함으로써 분리음원을 합성한다. 제안된 기법의 성능을 검증하기 위하여 SASSEC 에서 제공하는 테스트 음원 및 객관적 평가 지표를 이용하여 측정한 결과, 기존 ADRess 알고리즘에서 제시된 방법에 비해 개선된 품질의 분리음원을 합성하는 것으로 평가되었다.

1. 서론

스테레오 오디오 신호로부터 음원을 분리하는 기술은 객체 기반 오디오 서비스 또는 음악 정보 검색 서비스 등 다양한 응용 분야에서 활용될 수 있으며, 특히 현재 유통되는 대부분의 오디오 콘텐츠가 스테레오 채널로 제작되는 상황을 고려한다면 기술 활용도 측면에서 매우 큰 장점을 지닌다.

스튜디오에서 제작되는 오디오 콘텐츠는 보컬 및 다양한 악기 음원을 포함하고 있으며, 이는 분리해야 되는 음원의 개수가 입력되는 혼합 신호의 개수보다 많은 경우에 해당된다. 신호의 통계적 특성을 모델링하는 음원분리 방법은[1] 분리해야 될 음원의 개수보다 많은 혼합 신호를 필요로 하므로 입력되는 혼합 신호가 2 개인 스테레오 채널에 적용하기에는 적합하지 않다.

대표적인 스테레오 채널에 대한 음원분리 기술로 ADRess(Azimuth Discrimination and Resynthesis) 알고리즘이 있으며[2][3], 이는 인간의 좌/우 귀에 입력되는 오디오 신호 간 강도 차(IID: Inter-aural Intensity Difference)를 기반으로 음원의 위치를 인지하는 인간의 청각 특성을 이용한다. 대부분의 오디오 콘텐츠는 믹싱 단계에서 각 개별 음원에 대해 IID 기반의 패닝(panning)이 적용되므로, ADRess 알고리즘은 스테레오 채널에 대한 음원분리에 매우 유리하다.

ADRess 알고리즘에서는 매 프레임별로 추정된 방위각을 기준으로 인접한 frequency-azimuth 평면상의 값들을 더해 주파수 영역에서의 크기 성분을 구하고, 입력된 스테레오 신호와 동일한 위상 성분을 적용함으로써 분리음원의 합성 과정을 수행한다. 본 논문에서는 앞서 설명한 기존 합성 방법 대신 추정된 방위각에 상응하는 신호 강도비로 표현되는 확률밀도함수를 좌/우 신호 중 하나의 주 혼합 신호에 취함으로써 분리음원을 합성할 수 있는 새로운 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2 절에서는 ADRess 알고리즘에서 사용하는 음원 방위각 식별 및 분리음원 합성 방법에 대해 살펴본 후, 3 절에서는 본 논문에서 제안한 새로운 분리음원 합성 방법에 대해 설명한다. 4 절에서는 SASSEC(Stereo Audio Source Separation Evaluation Campaign) 테스트 음원 및 객관적 평가 지표를 이용한 분리음원 합성 기법에 대한 성능평가 결과를 보여주고, 마지막으로 5 절에서는 본 논문에 대한 결론을 맺는다.

2. ADRess 알고리즘

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

2.1 음원 방위각 식별

매 신호 분석 프레임에 대해 STFT(Short-Time Fourier Transform)을 취하고, 식(1)을 이용하여 $(N+1)$ x $(\beta+1)$ 배열의 frequency-azimuth 평면을 구성한다. N과 β는 각각 주파수 해상도와 방위각 해상도를 의미한다.

$$A_Z(k,m,i)=\begin{cases}|X_2(k,m)-g(i)X_1(k,m)| & if\ i\le\beta/2\\ |X_1(k,m)-g(i)X_2(k,m)| & if\ i>\beta/2\end{cases} \quad (1)$$

여기서, k 는 $0\le k\le N$ 를 만족하고, $X_1(k,m)$ 과 $X_2(k,m)$ 는 각각 좌측과 우측 채널의 m 번째 프레임에서의 k번째 주파수 성분을 나타낸다.

식(2)로 정의되는 좌/우 채널간 신호 강도비 $g(i)$는 0 과 1 사이의 값을 가지며, 이때 i는 $0 \le i \le \beta$, i와 β는 정수이다.

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

β 값이 커질수록 더욱 정확한 방위각 식별이 가능하나, 계산량은 증가하게 되므로 적절한 값으로 설정한다. 엄밀히 말하면 ADRess 알고리즘에서는 정확한 방위각을 계산하지 않으며, azimuth 축은 신호 강도비 함수에서의 i 값에 해당한다.

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.

그림 1 은 사람의 음성 신호(방위각 100°)에 대해 계산된 frequency-azimuth 평면을 나타낸다. 그림에서 보는 바와 같이 음성 신호인 관계로 4kHz 이하 주파수 성분이 우세하며, 방위각 축의 100(본 논문에서는 방위각을 음원의 위치가 좌측인 경우 0, 중앙은 90, 우측은 180 으로 표시함) 근처에서 에너지가 몰려 있음을 알 수 있다.

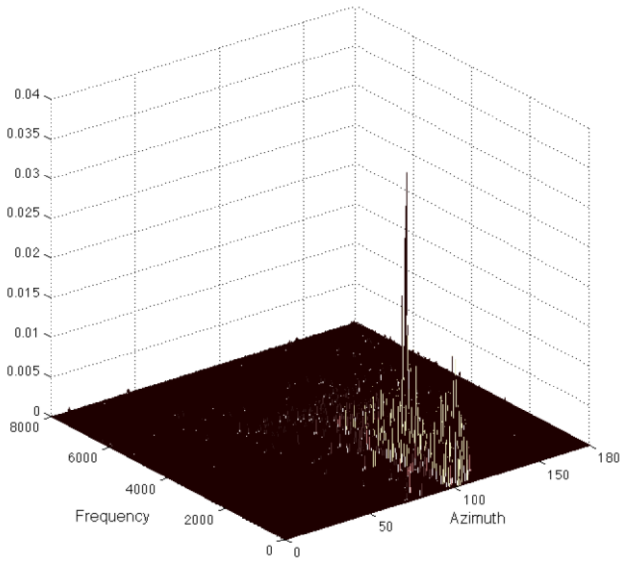

그림 1. Frequency-Azimuth 평면

2.2 분리음원 합성

그림 1 에서 보는 바와 같이 해당 음원의 에너지는 방위각을 중심으로 분포되어 있다. 따라서, j 번째 분리음원의 스펙트럼 크기는 식 (4)와 같이 식별된 방위각 인덱스 d_j 를 중심으로 방위각 서브공간 폭인 H 구간에 대해 $A_{\bar{Z}}(k,m,i)$ 를 누적해서 계산한다.

$$Y_j(k,m) = \sum_{i=d_j+H/2}^{i=d_j-H/2} A_{\bar{Z}}(k,m,i) \times \left(1 - \frac{2|d_j - i|}{H}\right) \quad (4)$$

상기 식의 두 번째 항은 선형 가중 함수로써, 삼각형 모양의 분리 윈도우를 나타낸다. H 값을 너무 크게 설정하면, 인접한 음원의 간섭으로 인한 왜곡 가능성이 커지므로 적정한 값을 선택해야 한다.

분리음원의 위상 정보는 식 (5)와 같이 입력되는 혼합 신호의 위상 정보를 구한 후, 식 (6)을 이용하여 주파수 영역에서의 최종 합성 신호 $S_j(k,m)$을 구성한다.

$$\Phi(k,m) = \angle\big(X(k,m)\big) \quad (5)$$

$$S_j(k,m) = Y_k(k,m) \cdot e^{j\Phi(k,m)} \quad (6)$$

ISTFT(Inverse-STFT)를 통해 구한 매 프레임별 시간 영역에서의 분리음원 신호는 overlap-add 기법에 의해 재결합됨으로써 분리음원 합성 과정을 마무리한다.

3. 제안된 분리음원 합성 기법

기존 ADRess 알고리즘에서는 frequency-azimuth 평면상의 azimuth 축을 신호 강도비의 인덱스 i를 기준으로 구성하므로 실제 방위각을 의미하진 않는다. 이에 sinusoidal energy-preserving panning law 를 기반으로 식(7)과 같은 실제 방위각과 신호 강도비간 관계를 유도한다.

$$azimuth(i) = \begin{cases} \dfrac{360° \cdot artan(g(i))}{\pi} & if\ i \le \beta/2 \\ 180° - \dfrac{360° \cdot artan(g(i))}{\pi} & if\ i > \beta/2 \end{cases} \quad (7)$$

여기서 i는 $0 \le i \le \beta$, i와 β는 정수이다.

먼저, 신호 강도비 $g(i)$ 는 기존 ADRess 알고리즘에서 정의한 식(2) 대신 식(7) 함수를 이용하여 각 신호 분석 프레임별 frequency-azimuth 평면을 구성한다. 구성된 평면에서 방위각에 따라 누적된 $A_{\bar{Z}}(k,m,i)$ 에너지 값을 이용하여, 분리하고자 하는 음원의 개수만큼 peak 값들을 찾음으로써 해당 peak 들이 위치하는 정확한 방위각 d_j를 구할 수 있다. 그림 2 는 4 개의 음원이 혼합된 스테레오 오디오 신호에 대한 음원 방위각 식별 결과로써, 원래 음원의 정확한 방위각을 찾아내는 것을 알 수 있다.

식(7)의 신호 강도비 $g(i)$는 방위각 90°를 중심으로 좌우 대칭값을 갖는 관계로 좌/우측 방위각에 대한 패닝 모호성을 갖는다. 이를 해결하도록 식(8)과 같이 새로운 $\bar{g}(i)$ 함수를 정의하며, 그림 3 은 방위각 변화에 따른 $\bar{g}(i)$값을 보여준다.

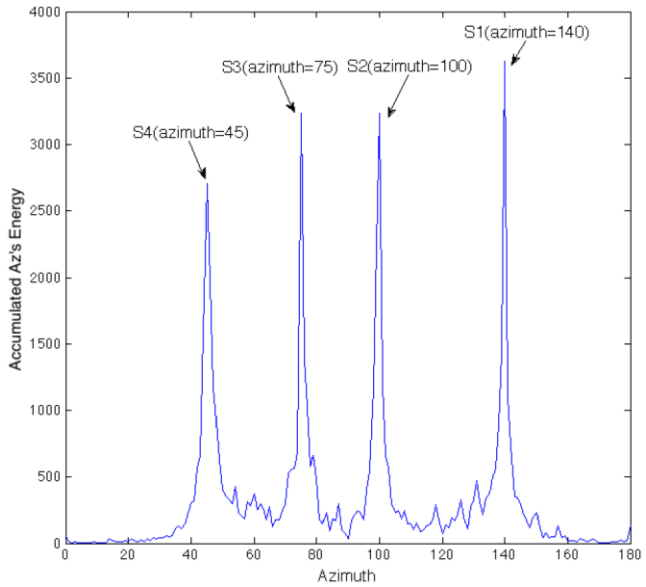

그림 2. 음원 방위각 식별 결과

$$\bar{g}(i) = \begin{cases} (1 - g(i)) \cdot (-1) & if\ i \le \beta/2 \\ 1 - g(i) & if\ i > \beta/2 \end{cases} \quad (8)$$

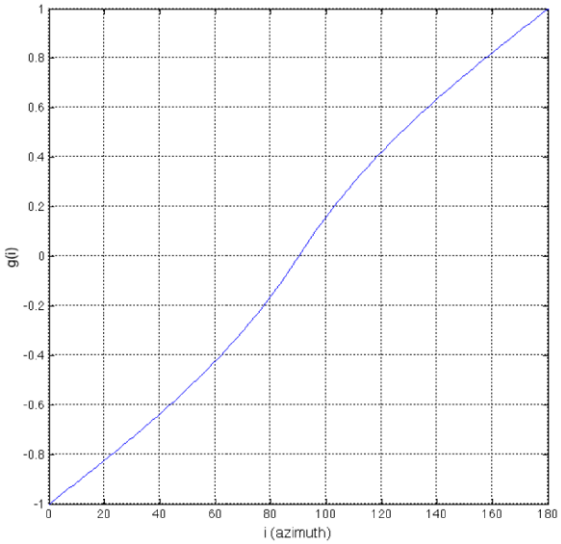

그림 3. 방위각 변화에 따른 $\bar{g}(i)$

분리음원 합성을 위해 식(10)와 같이 식별된 방위각에 상응하는 신호 강도비로 표현되는 확률밀도함수로 가우시안 윈도우 함수 $G_j(k,m)$를 정의한다.

$$U(k) = \arg\min_{0 \le i \le \beta} A_Z(k,m,i) \quad (9)$$

$$G_j(k,m) = \frac{1}{\sqrt{2\pi\gamma}} e^{-\left(\bar{g}(U(k)) - \bar{g}(d_j)\right)^2/(2\gamma)} \quad (10)$$

여기서, γ 는 윈도우 폭을 제어하며, 분리음원의 왜곡 성분이 커지지 않도록 적절히 설정되어야 한다. $U(k)$ 는 식(9)에서와 같이 매 신호 분석 프레임별로 k 번째 주파수 성분에서 $A_Z(k,m)_{min}$ 을 갖는 인덱스 i 값이다.

식(11)에서와 같이 앞서 정의한 가우시안 윈도우 함수를 좌/우 신호 중 하나의 주 혼합 신호에 취함으로써 분리음원에 대한 주파수 영역 신호를 구한다.

$$S_j(k,m) = \begin{cases} G_j(k,m) \cdot X_1(k,m) & if\ d_j \le \beta/2 \\ G_j(k,m) \cdot X_2(k,m) & if\ d_j > \beta/2 \end{cases} \quad (11)$$

ISTFT 를 취한 후, 시간 영역 신호에 대해 overlap-add 기법을 적용함으로써 최종적으로 분리음원을 생성한다.

4. 실험 결과

본 논문에서 제안한 분리음원 합성 기법의 성능을 평가하기 위해 SASSEC 에서 제공하는 테스트 음원 및 객관적 평가 지표를 이용하였다[4].

테스트 음원은 2 개의 무지향성 마이크로폰을 이용하여(이격 거리: 5cm) 4 개의 방위각(45º, 75º, 100º, 140º)에 대해 1m 반경으로 위치한 스피커마다 서로 다른 4 명의 여성 음성을 입력으로 하여 녹음되었다.

객관적 평가지표로는 식(13) ~ 식(15)로 표현되는 SDR(Source to Distortion Ratio), SIR(Source to Interference Ratio), SAR(Source to Artifact Ratio)을 사용하였으며, 이는 식(12)에서와 같이 추출된 분리음원 $\hat{s}(t)$에 대한 성분 분해를 통해 다음과 같이 정의된다[5].

$$\hat{s}(t) = s_{target}(t) + e_{interf}(t) + e_{noise}(t) + e_{artif}(t) \quad (12)$$

$$SDR = 10\log_{10} \frac{\|s_{target}\|^2}{\|e_{interf} + e_{noise} + e_{artif}\|^2} \quad (13)$$

$$SIR = 10\log_{10} \frac{\|s_{target}\|^2}{\|e_{interf}\|^2} \quad (14)$$

$$SAR = 10\log_{10} \frac{\|s_{target} + e_{interf} + e_{noise}\|^2}{\|e_{artif}\|^2} \quad (15)$$

기존 ADRess 알고리즘을 기반으로 방위각 서브공간 폭(H)을 3 으로 설정하였을 경우, 4 개의 분리음원에 대한 성능 분석 결과는 표 1 과 같다.

표 1. ADRess 기반 분리음원 합성 성능평가 결과($H = 3$ 기준)

	SDR	SIR	SAR
분리음원 S1	−1.0191	21.6249	−0.9656
분리음원 S2	−3.5087	16.8600	−3.3801
분리음원 S3	−4.3874	16.7978	−4.2643
분리음원 S4	−2.6588	21.0280	−2.6060

성능평가 결과, 분리음원 모두에서 SDR 및 SAR 이 낮은 값을 보이며, 이로 인해 합성 음질에서의 일부 열화가 발생하였다. 표 2 에서 보는 바와 같이 제안된 분리음원 합성 기법의 경우, SDR 및 SDR 이 개선되어 분리음원에 대한 음질이 향상되었다.

표 2. 제안 방식 기반 분리음원 합성 성능평가 결과

	SDR	SIR	SAR
분리음원 S1	5.1997	22.0674	5.3169
분리음원 S2	7.5229	19.8289	7.8310
분리음원 S3	5.6897	19.0912	5.9460
분리음원 S4	6.4391	21.0997	6.6238

시간 영역에서의 원음원과 분리음원 간 신호 파형은 그림 4 와 같으며, 원신호 파형에 맞춰 분리음원 합성이 적절히 이뤄졌음을 알 수 있다.

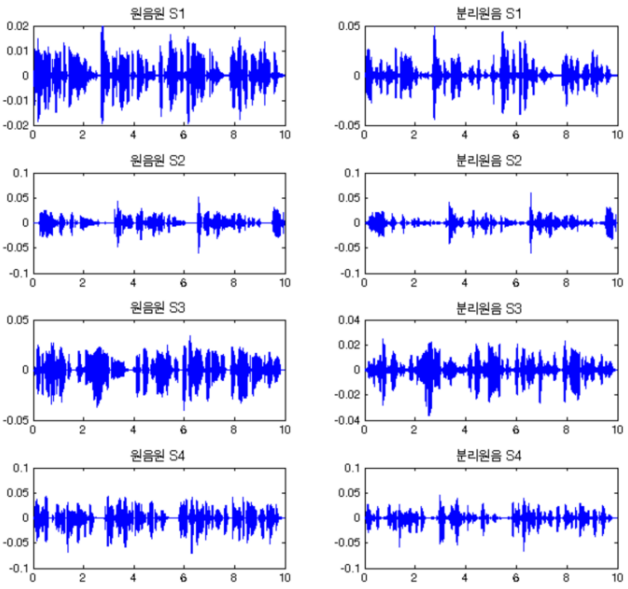

그림 4. 원음원과 분리음원 간 신호 파형 비교

표 3. 분리음원 합성 기법 간 성능평가 결과 비교

	SDR	SIR	SAR
기존 방식	−2.89	19.07	−2.80
제안 방식	6.21	20.52	6.43

표 3 은 제안된 분리음원 합성 기법과 기존 방식 간 성능평가 비교 결과를 나타낸다. 모든 평가지표에서 제안 방식이 기존 방식에 비해 개선된 결과를 나타냈으며, 평가 지표별로 살펴보면 SIR 은 1.45dB, SDR 은 9.1dB, SAR 은 9.23dB 향상되었다.

5. 결론

본 논문에서는 스테레오 오디오 신호로부터 추정된 음원 방위각에 대한 신호 강도비를 이용하여 분리음원을 생성하는 새로운 분리음원 합성 기법을 제안하였다. 이를 위해 신호 분석 프레임별로 개선된 신호 강도비 함수를 적용한 frequency-azimuth 평면을 구성하고, 이를 통해 추정된 방위각에 상응하는 신호 강도비로 표현되는 확률밀도함수를 좌/우 신호 중 하나의 주 입력 신호에 취함으로써 분리음원을 합성하였다. 제안된 분리음원 합성 기법에 대한 성능 평가는 SASSEC 에서 제공하는 테스트 음원 및 객관적 성능평가 방법을 이용하여 실시하였다. 기존 ADRess 알고리즘 기반 분리음원 합성 방법과의 성능평가 비교 결과, SIR 은 1.45dB, SDR 은 9.1dB, SAR 은 9.23dB 각각 개선된 품질의 분리음원을 합성하는 것으로 평가되었다.

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] A. Hyvärinen, “Survey on Independent Component Analysis,” Neural Computing Surveys 2, pp.94-128, 1999.

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

[3] J. Mckay et al., “Evaluating Ground Truth for ADRess as a Preprocess for Automatic Musical Instrument Identification,” AES 126th Convention, May. 2008.

[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.