비음수 행렬 분해 및 일반화된 상호상관계수 기법을 이용한 TV시청 환경에서의 다중 음원 방향 추정 방법

유승우, 전광명, 박지현, 김홍국

광주과학기술원

{yuseungwoo, kmjeon, jihyun, hongkook}@gist.ac.kr

Direction Estimation of Multiple Sound Sources Using Non-negative Matrix Factorization and Generalized Cross-Correlation

Seung Woo Yu, Kwang Myung Jeon, Ji Hyun Park, Hong Kook Kim

Gwangju Institute of Science and Technology (GIST)

요약

본 논문에서는 실내 환경 중 TV 시청환경에서 마이크로폰 어레이를 이용하여 다양한 다중 음원 방향을 추정하는 기법을 제안한다. 제안된 기법은 기존의 하나의 음원에 특화되어 있는 GCC-PHAT 기반의 방법을 GCC-PHAT 버퍼와 NMF를 도입하여 다중음원의 방향 추정을 가능하게 만들었다. 제안된 기법의 성능을 평가하기 위해서 실 거주 환경에서 발생하는 소음원과 TV 소리 방향 추정 결과에 대한 실측치와 추정치 간의 오차인 절대 평균오차를 측정하였으며, 실험 결과 제안한 기법이 기존의 방법인 GCC-PHAT보다 우수한 추정 성능을 보임을 확인하였다.

1. 서론

실내 환경에서 발생하는 소음은 업무, 독서, TV 시청시 집중력을 흐리게 하는 원인이다. 특히 TV 시청시 콘텐츠의 소리 이외의 주변의 소음은 잡음으로 간주되어 집중에 방해가 된다 [1]. TV 시청환경에서 원하는 콘텐츠 소리 이외에 소음원을 제거하고자 마이크로폰 배열을 사용하는 방법이 연구되고 있다 [1]. 마이크로폰 배열 기반 잡음제거 기법은 임의의 공간상에서 음원 방향을 탐지하여 그 방향의 음원을 추출한다. 그러므로 정확한 음원 방향 추청 기법의 연구가 필수적이다. 기존의 방법으로는 time delay of arrival (TDOA) 기반인 generalized cross-correlation phase transform (GCC-PHAT)의 공간 해상도를 보상한 phase error minimization (PEM) 기준을 사용해왔다 [2]. 하지만 기존 방법은 채널 별 상관도에 의존하여 위상의 파워가 큰 신호만을 찾게 되므로 하나의 음원에 편중되어 찾게 되는 현상이 발생하게 된다 [3]. 따라서, 이러한 단점을 극복하고자 GCC-PHAT 버퍼와 non-negative matrix factorization (NMF)을 이용하여 GCC-PHAT의 다중음원 방향 추정을 가능하게 하고자 한다.

본 논문의 구성은 다음과 같다. 2절에서는 NMF 기반의 GCC-PHAT을 이용한 음원 위치 추정 알고리즘을 제안한다. 그리고 3절에서는 제안된 방식의 성능을 평가한 후, 4절에서는 본 논문에 관한 결론을 맺는다.

2. 제안된 NMF GCC-PHAT 위치 추정 알고리즘

2.1 PEM 기준 GCC-PHAT

GCC-PHAT은 한 쌍의 마이크로폰 신호간의 상호 상관으로 정의되며, 푸리에 변환에 의한 주파수 영역에서 다음 식과 같이 표현될 수 있다 [2].

$$R_{l,n} = \sum_{\omega=-\omega_s}^{\omega_s} \Psi_n(\omega) X_{1,n}(\omega) X_{2,n}^*(\omega) e^{-iw\beta} \quad (1)$$

여기서, $R_{l,n}$는 n번째 프레임에 해당하는 $((2L+1)\times 1)$차원의 열벡터이다. l은 마이크 채널간 시간지연 샘플을 의미하며, L은 최대 시간지연 샘플이다. $\omega(=[-\omega_s,\omega_s])$는 radian으로 표현되는 주파수 영역의 인덱스이며, ω_s는 샘플링 주파수이다. $X_{1,n}(\omega)$는 n번째 프레임에서 1번 채널에서 녹음되어 주파수 영역에서 표현된 신호다. 또한, *는 conjugate를 의미하며, $\Psi_n(\omega)=1/\left|X_{1,n}(\omega)X_{2,n}^*(\omega)\right|$는 실내 잔향환경에 가장 좋은 성능을 보이는 PHAT 가중치이다. 또한, 식 (1)은 아래 식과 같이 표현될 수 있다 [2].

$$R_{l,n} = \sum_{\omega=-\omega_s}^{\omega_s} \cos(\theta_n(\omega)) \quad (2)$$

여기서, $\theta_n(\omega)$는 채널간 위상차를 의미하며 $\theta_n(\omega) = \angle X_{1,n}(\omega) - \angle X_{2,n}(\omega) - \omega\beta$로 정의된다. 식 (2)에서 $\theta_n(\omega)$을 최소화는 시간 지연, β를 선택함으로써 음원의 각도를 추정할 수 있다. 그리고 마이크간의 거리를 d, 음속을 c, 채널간의 위상지연을 $p(=[-90^\circ, 90^\circ])$라 하면, $\beta = d\sin(p)/c$가 된다 [2].

다중음원의 방향 추정을 가능케 하기 위해 $R_{l,n}$을 저장하는 버퍼, $G_{p,b}$를 두어 프레임의 연속성을 고려하면 $G_{p,b}=[R_{l,n-B},\cdots,R_{l,n}]$와 같이 표현될 수 있다. $b(=[n-B,n])$는 버퍼에 존재하는 프레임 인덱스를 나타내며, B는 버퍼의 길이를 의미하며 본 논문에서는 10으로 설정하였다.

2.2 NMF 기반 GCC-PHAT

주어진 GCC-PHAT 버퍼, $G_{p,b}$를 NMF로 다음과 같이 분해가 가능하다 [4].

$$G_{p,b} \simeq C_{p,k}E_{k,b} \tag{3}$$

여기서, $C_{p,k}$의 열은 NMF의해 분해된 기저행렬(basis)이라고 부르며 차원 k번째 음원에 따른 위상지연, p와의 관계를 의미한다. $E_{k,b}$의 열을 활성행렬(activation)이라고 부르고 b번째 프레임에서 가장 활동적인 k번째 음원을 나타낸다. 본 논문에서 K는 10으로 설정하였다. 식 (3)의 행렬 $C_{p,k}$와 $E_{k,b}$는 Kullback-Leibler (KL) divergence를 최소화하는 기준에 의해 다음과 같이 업데이트된다 [4].

$$E_{k,b} \leftarrow E_{k,b} \otimes \frac{C_{p,k}^T(G_{p,b} \otimes (C_{p,k}E_{k,b})^{-1})}{C_{p,k}^T} \tag{4}$$

$$C_{p,k} \leftarrow C_{p,k} \otimes \frac{((C_{p,k}E_{k,b})^{-1} \otimes G_{p,b})E_{k,b}^T}{E_{k,b}^T} \tag{5}$$

활성행렬과 $G_{p,b}$를 이용하여 식 (2) 대신 k번째 음원별 PEM 기준으로 GCC-PHAT 버퍼를 예측할 수 있다 [5].

$$\hat{G}_{p,k} = \frac{\sum_{n \in N_k} G_{p,b}E_{k,b}^T}{\sum_{n \in N_k} E_{k,b}^T} \tag{7}$$

여기서, N_k는 활성행렬에서 k번째 음원이 가장 높은 활성도를 나타낸 k음원별 프레임 인덱스를 의미한다 [5].

2.3. 음원 방향 추정

음원 방향은 식 (7)의 버퍼를 최대화함으로써 추정할 수 있다. 즉,

$$\hat{\theta}_{c,k} = argmax_{\theta \in p}(\hat{G}_{p,k}) \tag{8}$$

다중 음원을 추정할 경우 식(8)의 과정을 반복한 후, 복수개의 각도를 찾을 수 있다. 다만, 식 (8)을 반복적으로 수행 시, 이전 반복에서 얻은 각도는 추정범위에서 제외한다.

3. 성능평가

제안된 방법의 성능을 평가하기 위해서 다음과 같이 실험환경을 구성하였다. 음원은 약$112.39m^2$의 실내 환경에서 녹음되었다. 1개 음원(src1)은 19종류 9분 분량의 음원으로 이루어져 있으며, 2개 음원은 src1과 TV로 구성되었다. 이때, TV는 광고, 드라마, 뉴스, 스포츠, 예능으로 세분화되어 있다. 각 음원은 16-bit의 해상도를 갖고 48 kHz 샘플링 주파수로 녹음되었으며, 한 프레임의 사이즈는 80ms이며 중첩사이즈는 40ms이었다. 비교 방법으로는 식 (2)의 PEM 기준의 GCC-PHAT 도래각 추정 방법을 선정하였다.

음원의 측정치와 추정치의 차이의 절대 오차 평균인 mean absolute error (MAE) 사용하여 다음 식과 같이 정략적 평가를 하였다.

$$\epsilon_{DOA} = \frac{1}{I}\sum_{i=1}^{I} |\theta_{ref}(i) - \theta_{esti}(i)| \tag{9}$$

여기서, $\theta_{ref}(i)$와 $\theta_{esti}(i)$는 원음의 각도 및 추정된 음원의 각도이다. 화각을 고려한 임계치를 설정하였고 오차가 임계치 이내면, 유효 프레임이라고 결정하였다. 즉, 식 (9)에서 I는 추정된 유효 프레임수이다. 본 실험에서는 2개 음원방향 추정시 TV(55°) 기준 최소 12°, 최대 80°에 떨어져 있으므로 임계치는 10°로 정하였다.

또한, 프레임 추정 평균 오차는 $\epsilon_{det} = (N-I)/N$로 표현되며, 이때 N은 src1 종류에 따른 프레임 총수를 의미한다.

표 1. 방향추정 평균 오차 (ϵ_{DOA}) 비교

음원 종류		GCC-PHAT	제안된 방법
1개 음원	src1	4.26	3.76
2개 음원	TV(55°)	4.90	5.09
	src1	4.25	3.71

표 2. 프레임 추정 평균 오차 (ϵ_{det}) 비교

음원 종류		GCC-PHAT	제안된 방법
1개 음원	src1	0.22	0.16
2개 음원	TV(55°)	0.57	0.55
	src1	0.33	0.28

4. 결론

본 논문에서는 TV 시청 환경에서 소음원 방향 추정기술에 대해서 NMF 기반의 GCC-PHAT 다중 소음원 추정 기법을 제안하였다. 제안된 방법은 GCC-PHAT 방향 추정 대비 낮은 오차를 보이며, 프레임 추정 평균 오차를 최대 6% 정도 감소시켰다.

감사의 글

본 연구는 미래창조과학부 및 정보통신기술연구진흥센터의 정보통신·방송 연구개발사업의 일환으로 수행하였음 [B0101-15-1360, 라우드니스 기반의 기술 및 실내 황경 소음의 스트레스 평가 기술 개발].

참고 문헌

[1]. 전광명, 이동윤, 유승우, 김홍국, “TV 시청환경에서의 실내소음의 스트레스 평가를 위한 소음 데이터베이스 설계,” *2014년 한국방송공학회 추계학술대회,* paper no. P2-14, 2014.

[2]. M. F. Font, *Multi-microphone Signal Processing for Automatic Speech Recognition in Meeting Rooms*, MS Thesis, Universitat Politecnica de Catalunya, Spain, 2005.

[3]. H. Do and H. F. Silverman, “Robust cross-correlation-based techniques for detecting and locating simultaneous, multiple sound sources,” in *Proc. ICASSP*, pp. 201-204, 2012.

[4]. J. Le Roux, F. J. Weninger, and J. R. Hershey, *Sparse NMF - Half-Baked or Well Done?*, Tech. Rep. TR2015-023, MERL, Cambridge, MA, Mar. 2015.

[5]. H. Kayser, H. Anemuller, and K. Adiloglu, “Estimation of inter-channel phase differences using non-negative matrix factorization,” in *Proc. IEEE Sensor Array and Multichannel Signal Processing Workshop*, pp. 77-80, 2014.