휴대용 단말기에서 음원 위치 추적 기술 비교 연구

정 재 연, 육 동 석*
고려대학교 컴퓨터학과

A Comparative Study of Sound Source Localization Algorithms for Portable Devices

Jaeyoun Chung, Dongsuk Yook
Department of Computer Science and Engineering, Korea University
E-mail : tochan@voice.korea.ac.kr, yook@voice.korea.ac.kr

Abstract

The performance of a sound source localization system degrades severely in reverberant and noisy environments. In addition, restriction on the distance between microphones, which is required by portable devices, also lower the system performance. This paper compares the sound source localization algorithms based on time delay of arrival, which are robust to reverberation and noises considering microphone sensor distance. As well, post filter which outputs maximum count time delay is adopted to increase the accuracy.

I. 서론

최근 여러 개의 마이크로폰을 이용하여 음성을 처리하는 연구가 많이 이루어지고 있다[1]. 그 중 음원의 위치를 추적하는 기술은 화자의 위치 정보를 이용하여 화자에게 자동으로 초점을 맞추는 화상 회의 시스템[2]이나 화자의 위치 정보를 이용하여 강화된 음성을 사용하는 음성 인식 시스템[3] 등 다양한 분야에 적용되고 있다. 특히 핸드폰, PDA 등의 휴대용 단말기에서의 음성 인식 시스템[4]이나 청각 장애인 혹은 휴머노이드 로봇에게 필요한 인공청각 시스템[5][6] 등에 활용될 수 있기 때문에 중요한 기술이라고 할 수 있다.

음원 위치 추적 기술은 일반적으로 세 가지 종류로 분류할 수 있다. 첫 번째는 조향된 빔형성기(steered beamformer)를 이용한 방법, 두 번째는 고해상도 스펙트럼 추정(high resolution spectral estimation)을 이용한 방법, 그리고 세 번째는 도착 지연 시간(time delay of arrival)을 이용한 방법이다[7]. 이 중 세 번째인 도착 지연 시간을 이용한 방법이 다른 방법들에 비해 계산의 복잡도가 낮고 정확성이 높기 때문에 가장 많이 연구되고 있다[7]. 대부분의 방법들이 반향과 잡음이 없는 환경에서는 정확한 결과를 보여주지만 반향과 잡음이 있는 실제 환경에 적용한다면 정확성이 현저히 떨어진다. 특히 휴대용 단말기, 휴머노이드 로봇, 착용형 컴퓨터 등은 움직이는 장치이기 때문에 주변 잡음뿐 아니라 기기내 잡음을 포함한 낮은 SNR(signal-to-noise ratio) 환경과 실내에서의 반향 환경을 반드시 고려해야 한다[6]. 그리고 휴대용 단말기, 휴머노이드 로봇, 착용형 컴퓨터 등에 여러 개의 마이크로폰을 장착할 경우 장착할 수 있는 공간에 제약이 있기 때문에 마이크로폰 사이의 간격이 제한적임을 고려해야 한다. 본 논문에서는 마이크로폰 사이의 간격이 좁은 상황에서 도착 지연 시간 방법을 기반으로 반향 효과 감소 및 잡음 제거를 위한 음원 위치 추적 기술들을 비교하고, 정확성을 높이기 위해 빈도가 가장 높은 지연 시간을 출력하는 후처리 필터를 적용하여 그 결과를 비교한다.

---

* 교신저자
본 연구는 한국과학재단 특정기초연구 (R01-2006- 000-11162-0) 지원으로 수행되었음.

본 논문은 다음과 같이 구성된다. 2장에서는 신호를 모델링하여 도착 지연 시간을 이용한 일반적인 음원 위치 추적 방법, 반향 효과 감소 방법, 잡음 제거 방법, 그리고 후처리 필터 방법을 다루고, 3장에서는 2장에서의 방법들을 이용하여 실험한 결과들을 비교하고, 4장에서 결론을 맺는다.

II. 도착 지연 시간 추정 방법

2.1 신호 모델링

$x_1(n)$과 $x_2(n)$을 각각 첫 번째와 두 번째 마이크로폰에서 받은 신호라고 하면 반향, 잡음, 그리고 지연 시간을 고려한 신호 모델링은 다음과 같다.

$$x_1(n) = h_1(n) * s(n) + n_1(n) \qquad (1)$$
$$x_2(n) = h_2(n) * s(n-D) + n_2(n)$$

D은 구하고자 하는 지연 시간, $h_1(n)$과 $h_2(n)$는 반향에 해당하는 충격 응답(impulse response), $s(n)$은 원본 음원 신호, $n_1(n)$과 $n_2(n)$는 각 마이크로폰에서의 무상관 잡음(uncorrelated noise), 그리고 $*$는 선형 승적(linear convolution)을 나타낸다.

2.2 GCC(Generalized Cross-Correlation) 기반의 지연 시간 추정 및 반향 효과 감소 방법

$s(n)$과 $s(n-D)$간의 교차상관(cross correlation)인 $R_{ss}(\tau)$를 이용하거나 이를 푸리에 변환(the Fourier transform)한 파워스펙트럼(power spectrum)인 $G_{ss}(\omega)$를 이용하여 지연시간을 추정할 수 있다. 그렇지만 $R_{ss}(\tau)$나 $G_{ss}(\omega)$를 직접 구할 수 없기 때문에 마이크로폰에서 받은 신호 $x_1(n)$과 $x_2(n)$을 이용한다. 신호와 잡음 간에 상관관계가 없다고 가정한다면, 식 (1)을 이용한 $x_1(n)$과 $x_2(n)$의 파워스펙트럼은 다음과 같다.

$$G_{x_1x_2}(\omega) = H_1(\omega)H_2^*(\omega)G_{ss}(\omega)e^{-j\omega D} + G_{n_1n_2}(\omega) \qquad (2)$$

$G_{x_1x_2}(\omega)$는 $x_1(n)$과 $x_2(n)$의 파워스펙트럼, $H_1(\omega)$과 $H_2(\omega)$는 각각 $h_1(n)$과 $h_2(n)$의 푸리에 변환, *는 켤레 복소수(complex conjugate)변환, 그리고 $G_{n_1n_2}(\omega)$는 $n_1(n)$과 $n_2(n)$의 파워스펙트럼을 나타낸다. 또한 잡음 $n_1(n)$과 $n_2(n)$ 간에 상관관계가 없다면, 식 (2)는 다음과 같이 표현된다.

$$G_{x_1x_2}(\omega) = H_1(\omega)H_2^*(\omega)G_{ss}(\omega)e^{-j\omega D} \qquad (3)$$

식 (3)을 이용하여 GCC 기반으로 지연 시간을 추정하는 식은 다음과 같다[8].

$$D = \arg\max_{\tau} R_{x_1x_2}(\tau)$$
$$R_{x_1x_2}(\tau) \approx \frac{1}{2\Pi}\int_{-\Pi}^{\Pi} W(\omega)G_{x_1x_2}(\omega)e^{j\omega\tau}d\omega \qquad (4)$$

$W(\omega)$는 주파수 가중함수로서 식 (2)와 (3)에 나타난 $H_1(\omega)H_2^*(\omega)$의 역수이다. $h_1(n)$과 $h_2(n)$를 반향에 해당하는 충격 응답으로 모델링했기 때문에 $W(\omega)$을 적절하게 설정하면 반향 효과를 감소할 수 있다. 이를 위해 다양한 가중 함수들이 제안되어 왔다[2][8]. 그 중 입력 신호를 백색화(whitening)하는 PHAT(Phase Transform) 방법의 가중 함수는 다음과 같다.

$$W_{PHAT}(\omega) = \frac{1}{|G_{x_1x_2}|} \qquad (5)$$

PHAT 방법은 다른 방법과는 달리 원본 신호와 잡음간의 통계적인 특성을 모델링할 필요가 없고, 각각의 주파수 구간들에 동일한 비중을 두기 때문에 반향의 영향을 감소하는 효과가 있다. PHAT를 가중함수로 사용하는 GCC 방법을 GCC-PHAT라고 한다.

2.3 잡음 제거 방법

소리가 나지 않은 구간에서 잡음의 특성을 측정하고, 이를 이용하여 잡음을 제거한다.

(1) *스펙트럼 차감법(Spectral Subtraction: SS)*[9]

스펙트럼 차감법은 잡음이 포함된 신호에서 잡음의 스펙트럼 성분을 감하여 잡음이 제거된 원본 신호를 얻는 방법이다.

$$Y_i(\omega) = X_i(\omega)\left(\sqrt{1 - \frac{|N_i(\omega)|^2}{|X_i(\omega)|^2}}\right) \quad i = 1,2 \qquad (6)$$

(2) *G_{nn} 차감법(G_{nn} Subtraction: GS)*[2]

식 (2)에서 식 (3)으로 변환할 때 $n_1(n)$과 $n_2(n)$ 간에 상관관계가 없다고 가정하였지만, 실제 환경에서는 상관관계가 존재한다. G_{nn} 차감법은 잡음 간의 상관관

계를 고려하여 측정된 신호의 파워스펙트럼에서 잡음의 파워스펙트럼을 감하는 방법이다.

$$G_{s_1 s_2}^{GS}(\omega) = G_{x_1 x_2}(\omega) - G_{n_1 n_2}(\omega) \qquad (7)$$

2.4 높은 빈도 출력 후처리 필터

음원 위치 추적의 정확성을 높이기 위해 도착 지연 시간을 추정한 결과에 표 1의 높은 빈도 출력 후처리 필터(post filter)를 적용한다.

표 1 . 높은 빈도 출력 후처리 필터 처리 과정

1. 정해진 방법으로 현재 프레임의 도착 지연 시간을 추정한다. 2. 현재 프레임부터 필터 크기만큼 이전 프레임까지의 지연 시간을 필터 버퍼에 저장한다. 이 때 각각 프레임의 음성 구간 여부를 판단하고, 음성인 경우만 버퍼에 저장한다. 3. 필터 버퍼에 가장 빈도가 높은 지연 시간을 현재 프레임의 도착 지연 시간으로 출력한다.

III. 실험 결과

반향과 잡음이 존재하는 실내 환경에서의 가상 데이터를 생성하여 2장의 방법들로 실험하였다.

방의 크기는 7m×5m×3m로 설정하였고, 두 개의 마이크로폰은 바닥에서 1m 그리고 7m벽에서 1m 떨어진 가운데에 배치하였다. 마이크로폰 사이의 간격은 0.086m, 0.18m, 그리고 0.31m로 설정하였고, 음원은 마이크로폰과 약 20°, 50°, 그리고 80° 정도의 각도로 3m의 거리를 이루도록 설정하였다. 음원은 전화벨 소리를 사용하여 전체 30초의 길이에 15초의 소리 구간을 갖도록 구성하였다. 방의 반향은 반향 시간이 0.01초, 0.1초, 0.2초, 0.3초, 0.4초, 0.5초, 그리고 0.6초(반사 계수는 각각 0.002, 0.55, 0.74, 0.82, 0.86, 0.88, 그리고 0.90)가 되도록 설정하였고, 이미지 방법(image method)[10]을 사용하여 생성하였다. 잡음은 가우시안 잡음을 생성하여 SNR이 30dB가 되도록 넣었다. 샘플링 주파수는 48KHz로, 그리고 프레임의 크기는 1024 샘플(약 21ms)로 두고, 반 프레임씩 중첩되도록 하였다. 전체 음원은 2812개의 프레임이고, 그 중 실제 소리는 1406개의 프레임이다. 음원 위치 추적의 정확도는 측정하고자 하는 방향의 ±10° 이내인 경우를 맞은 것으로 하여 측정하였다. 표에서 정확도를 나타낸 단위는 모두 %이다.

표 2는 다양한 반향시간을 갖는 실험 데이터를 GCC (가중 함수를 1로 둔 방법)와 GCC-PHAT 방법으로 약 20°, 50°, 그리고 80°의 각도에 대해 평균 정확도를 측정한 표이다.

표 2. 추정 기법과 반향시간에 따른 정확도 비교
(마이크 간격: 0.086m, SNR: 30dB)

반향시간(초) / 추정기법	0.01	0.1	0.2	0.3	0.4	0.5	0.6
GCC	100	69.5	36.1	29.4	23.9	17.1	12.0
GCC-PHAT	77.7	50.0	36.3	30.2	27.8	24.5	20.0

반향 시간이 0.2초 미만일 때는 GCC 방법이, 0.2초 이상일 때는 GCC-PHAT 방법이 더 정확하였다. 이를 통해 반향이 심한 환경에서 GCC-PHAT가 GCC보다 높은 성능을 보임을 알 수 있지만, 정확도가 낮기 때문에 실제 상황에 사용하기는 어렵다.

표 3. 각도와 마이크로폰 간격에 따른 정확도 비교
(가) 추정기법: GCC, SNR: 30dB,
반향 시간: 0.1초, 후처리 필터 사용하지 않음

거리(m) / 각도	0.086	0.18	0.31
23°/17°/18°	12.1	50.0	53.1
54°/50°/49°	96.3	98.0	98.4
80°/80°/80°	100	100	99.5
평균	69.5	82.7	83.8

(나) 추정기법: GCC-PHAT, SNR: 30dB,
반향시간: 0.2초, 후처리 필터 사용하지 않음

거리(m) / 각도	0.086	0.18	0.31
23°/17°/18°	10.9	16.8	27.2
54°/50°/49°	34.2	32.2	35.7
80°/80°/80°	63.8	72.9	71.8
평균	36.3	40.6	44.9

(다) 추정기법: GCC-PHAT, SNR: 30dB,
반향시간: 0.2초, 0.4초 크기의 후처리 필터 사용

거리(m) / 각도	0.086	0.18	0.31
23°/17°/18°	20.4	37.5	66.2
54°/50°/49°	48.8	40.7	78.2
80°/80°/80°	96.5	97.2	95.0
평균	55.2	58.5	79.8

표 3은 마이크로폰 사이의 간격에 따른 정확도 변화를 측정한 표이다. 표 2에서 반향 시간이 0.2 미만일 때 GCC가 좋은 성능을, 반향 시간이 0.2 이상일 때 GCC-PHAT가 좋은 성능을 보였기 때문에 표 3의 (가)에는 30dB의 SNR과 0.1초의 반향 시간으로 GCC를 사용하였고, 표 3의 (나)에는 30dB의 SNR과 0.2초의 반향 시간으로 GCC-PHAT를 사용하였다. 표 3의 (다)에는 표 3의 (나)의 실험에 높은 빈도 출력 후처리 필터를 적용하였다. 표 3의 (가), (나), (다) 모두 마이크로폰 사이의 간격을 0.086m, 0.18m, 0.31m로 변화시키면서 0.086m일 때 음원을 23°, 54°, 80°의 각도로, 0.18m일 때 17°, 50°, 80°의 각도로, 그리고 0.31m일 때 18°, 49°, 80°의 각도로 배치하여 정확도를 측정하였다. 그 결과 표 3의 (가), (나), (다) 모두 마이크로폰 사이의 간격이 좁아질수록, 음원의 각도가 줄어들수록 정확도가 낮아졌다, 또한 반향의 증가로 인해 표 3의 (나)의 정확도가 표 3의 (가)의 정확도보다 낮아졌고, 후처리 필터의 사용으로 표 3의 (다)의 정확도가 표 3의 (나)의 정확도보다 높아졌다. 정확도의 변화가 심한 경우를 살펴보면, 표 3의 (가)에서는 49° 미만에서, 표 3의 (나)에서는 80° 미만에서 정확도가 많이 낮아졌으며, 특히 표 3의 (가)와 (나)에서 0.086m, 23°일 때 정확도가 각각 12.1%와 10.9%로 심각하게 낮아졌다. 표 3의 실험 데이터에 정확도의 허용치를 엄격히 적용하여 오차범위를 ±0°로 둔 결과는 각각 표 3의 (나)에서 0.086m 거리의 경우 평균 12.4%, 표 3의 (다)에서 0.086m 거리의 경우 평균 14.6%를 보였기 때문에 정확한 위치 정보를 요구하는 응용 분야에는 적용하기 어렵다.

Ⅳ. 결론

본 논문에서는 휴대용 단말기 같은 마이크로폰 사이의 간격에 제한이 있는 상황에서 도착 지연 시간 방법을 기반으로 반향 효과 감소 및 잡음 제거를 위한 음원 위치 추적 기술들을 비교 분석하였다. 반향이 작은 경우는 GCC가, 반향이 큰 경우는 GCC-PHAT가 보다 높은 정확도를 보였고, 마이크로폰 사이의 간격이 좁아질수록, 음원의 각도가 줄어들수록 보다 낮은 정확도를 보였다. 그리고 정확도를 높이기 위한 높은 빈도 출력 후처리 필터를 적용함으로 정확도가 향상됨을 보였다. 그렇지만 정확도의 허용치를 보다 엄격히 적용한 결과를 통해 마이크로폰 사이의 간격이 좁고 반향이 큰 경우 실제로 적용하기에는 정확도가 너무 낮음을 알 수 있었다. 따라서 이를 개선할 새로운 방법에 대한 연구가 필요하다.

참고문헌

[1] J.L. Flanagan and H.F. Silverman, *Workshop on Microphone Arrays: Theory, Design & Application.* New Brunswick, NJ: CAIP Ctr., Rutgers Univ. Press, 1994.

[2] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proc. ICASSP*, vol. 1, pp. 187-190, 1997.

[3] D. Giuliani, M. Omologo, and P. Svaizer, "Talker localization and speech recognition using a microphone array and a cross-powerspectrum phase analysis," in *Proc. ICSLP*, vol. 22-1, pp. 1243-1246, 1994.

[4] D. Giuliani, M. Matassoni, M. Omologo, and P. Svaizer, "Hands free continuous speech recognition in noisy environment using a four microphone array," in *Proc. ICASSP*, vol. 2, pp. 860-863, 1995.

[5] J.E. Greenberg and P.M. Zurek, "Evaluation of an adaptive beamforming method for hearing aids," *J. Acoust. Soc. Amer.*, vol. 91, no. 3, pp. 1662-1676, 1992.

[6] E. Mumolo, M. Nolich, and G. Vercelli, "Algorithms for acoustic localization based on microphone array in service robotics," *J. Robot. Autom.*, vol 42, no 2, pp. 69-88, 2003.

[7] M. Brandstein and H. Silverman, "A practical methodology for speech source localization with microphone arrays," *Comput., Speech Lang.*, vol. 11, no. 2, pp. 91-126, 1997.

[8] C. H. Knapp and C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Tran. Acoust., Speech, Signal Processing*, vol. ASSP-24, no. 4, pp. 320-327, 1976.

[9] S.F Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Tran. Acoust., Speech, Signal Processing*, vol. ASSP-27, no. 2, pp. 113-120, 1979.

[10] P. M. Peterson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," *J. Acoust. Soc. Amer.*, vol. 80, no. 5, pp. 1527-1529, 1986.