음성 신호를 이용한 시간지연 추정에 미치는 영향들에 관한 연구
Factors for Speech Signal Time Delay Estimation

권병호† ·박영진*·박윤식**
Byoungho Kwon, Youngjin Park and Youn-sik Park

Key Words : time delay estimation (시간지연 추정), the generalized cross correlation method (일반화된 상호상관 법), reverberation (잔향), R/D ratio (잔향-직접 음 에너지 비, reverberant to direct sound energy ratio).

ABSTRACT

Researches for time delay estimation had been studied broadly. However studies about factors for time delay estimation are insufficient, especially in case of real environment application. In 1997, Brandstein and Siverman announced that performance of time delay estimation deteriorates as reverberant time of room increases. Even though reverberant time of room is same, performance of estimation is different as the specific part of signals. In order to know that reason, we studied and analyzed the factors for time delay estimation using speech signal and room impulse response. In result, we can know that performance of time delay estimation is changed by different R/D ratio and signal characteristics in spite of same reverberant time.

요 약

시간지연 추정 방법들에 대한 연구는 예전부터 활발히 진행되고 있다. 하지만 실제 환경에서 측정된 신호들을 이용하여 시간지연을 추정함에 있어 그 성능에 미치는 영향들에 대한 연구는 미흡한 실정이다. 1997 년에 Brandstein 과 Siverman 은 공간의 잔향 시간이 길어질수록 그 공간에서의 시간지연 추정 성능이 나빠짐을 모의실험을 통해 밝혔다. 하지만 동일한 잔향시간을 갖는 공간에서 측정된 신호의 경우에도 시간 구간에 따라 추정 성능에 차이를 보이고 있다. 따라서 본 연구에서는 시간지연 추정 성능에 미치는 영향들에 대해서 살펴보고, 그 원인들에 대해 분석하였다. 그 결과, 동일한 잔향시간을 갖는 공간에서 측정된 신호일지라도 시간 구간에 따라 R/D 비와 신호의 특성들이 다르기 때문에 추정 성능에 차이가 나타남을 알 수 있었다.

1. 서 론

도달시간지연(Time Delay of Arrival, TDOA)을 이용한 음원 위치 추정 방법은 마이크로폰 어레이에 음향 신호가 도달되는 시간지연 값으로부터 음원의 위치를 추정하는 방법으로[1] 로봇의 청각 시스템[2], 자동 원격 회의를 위한 기술[3]등 많은 분야에 적용되고 있다. 이 방법에서 가장 중요한 부분은 마이크로폰 어레이에서 측정된 신호를 이용해 시간지연 값을 추정하는 것이다. 시간지연을 추정하는 방법들은 예전부터 많이 연구되어 왔는데, 1976 년에 Knapp 와 Carter 가 제시한 주파수 영역에서의 가중치 함수를 적용해 상호상관값(cross correlation)을 계산하고 이로부터 시간지연을 추정하는 방법이 가장 널리 사용되고 있다.[4] 뿐만 아니라 최소자승법을 이용하는 방법, Hilbert Trans form 을 이용하는 방법, Wavelet Transform 을 이용하는 방법 등도 연구되었다.[5-7]

앞서 언급한 측정된 신호로부터 시간지연 값을 구하는 방법들 뿐만 아니라, 마이크로폰이 존재하는 공간의 특성에 따른 시간지연 추정 성능에 대한 연구도 진행되었는데, 1997 년에 Brandstein 과 Silverman 은 공간의 잔향 시간이 길어질수록 그 공간에서의 시간지연 추정 성능이 나빠짐을 모의실험을 통해 밝혀 놓은바 있다.[8]

하지만 동일한 잔향 시간을 갖는 공간에서 측정된 신호를 이용해 시간지연 값을 추정할 때, 신호의 특정 부분에 따라 그 추정 성능에 차이가 발생함을 확인하였다. 이는 시간지연 추정 성능에 미치는 영향이 공간의 잔향 시간 외에도 다른 요인이 존재함을 의미한다. 따라서 본 연구에서는 시간지연 추정 성능에 영향을 미치는 영향들에 대해 공간의 잔향 특성과 측정된 신호의 특성의 관점에서 살펴보고자 한다. 먼저 본 연구에서 적용하고자 하는 시간지연 추정 방법은 가장 널리 사용되고 있는 참고문헌 (4)의 PHAT(Phase Transform) 가중치 함수를 적용한 경우로 제한

† KAIST 기계공학과
E-mail : bhkwon@kaist.ac.kr
Tel : (042) 869-3060, Fax : (042) 869-8220
* KAIST 기계공학과
** KAIST 기계공학과

한다. 그리고 공간의 잔향 특성에 따른 시간지연 추정 성능은 측정된 신호에 직접 음(direct wave) 에너지와 잔향 음(reverberation wave) 에너지 비를 나타내는 *R/D* 비를 이용하고, 측정된 신호의 특성은 그 신호의 주파수 특성과 신호의 크기변화에 따라 나타나는 시간지연 추정 성능에 대해 살펴보도록 한다.

2. 시간지연 추정 방법

시간지연 추정 성능에 미치는 영향들에 대해 알아보기 전에 본 연구에서 적용하고 있는 시간지연 추정 방법에 대해 간략하게 알아본다.

음원에서 방사된 음파가 두 마이크로폰에서 측정된 신호는 다음과 같이 표현될 수 있다.

$$x_1(t) = s(t) + n_1(t), \qquad \text{식(1)}$$

$$x_2(t) = \alpha s(t-D) + n_2(t). \qquad \text{식(2)}$$

x_i 는 i 번째 마이크로폰에서 측정된 신호이고, $s(t)$ 는 음원, $n_i(t)$ 는 i 번째 마이크로폰에서 측정된 잡음 신호를 나타낸다. $s(t)$, $n_1(t)$, $n_2(t)$ 는 real, jointly stationary random process 이고 $s(t)$ 는 $n_1(t)$, $n_2(t)$ 와 비연관 되어있다고 가정한다. 이 때 α 는 감쇄계수이고, D 는 추정하고자 하는 시간지연이 된다. 두 신호 x_1 , x_2 의 상호상관값은 다음과 같이 계산할 수 있다.

$$R_{x_1x_2}(\tau) = E[x_1(t)x_2(t-\tau)]. \qquad \text{식(3)}$$

이 때 마이크로폰 사이의 시간지연은 식(3)에서 $R_{x_1x_2}(\tau)$ 가 최대값을 갖게 하는 τ 가 된다. 그러나 일반적으로 마이크로폰으로부터 측정된 신호 x_1 , x_2 는 유한한 길이를 갖기 때문에 $E[\bullet]$ 을 이용해서 $R_{x_1x_2}(\tau)$ 을 정확하게 구할 수 없다. 따라서 유한한 길이를 갖는 신호들을 ergodic processes 라고 가정하고 식(4)로부터 상호상관값을 추정해야 한다.

$$\hat{R}_{x_1x_2}(\tau) = \frac{1}{T-\tau}\int_{\tau}^{T} x_1(t)x_2(t-\tau)dt. \qquad \text{식(4)}$$

T 는 측정된 신호의 길이이다. 또한 상호상관값, $R_{x_1x_2}(\tau)$ 는 상호스펙트럼밀도함수, $G_{x_1x_2}(f)$ 와 퓨리에 변환 관계가 있음을 알고 있다. 이와 같은 성질을 이용하면서, 식(4)에서 추정된 값과 실제 값 사이의 오차를 줄이기 위해서, 식(5)와 같이 상호스펙트럼밀도함수에 가중치 함수, $\psi_g(f)$ 을 곱해 좀 더 정확한 상호상관값을 추정하고, 이로부터 시간지연을 구하는 방법이 일반화된 상호상관값(generalized cross correlation) 방법이다.[4] 여러 가중치 함수 중에 PHAT 가중치 함수는 식(6)과 같이 표현된다.

$$\hat{R}^{(g)}_{x_1x_2}(\tau) = \int_{-\infty}^{\infty} \psi_g(f)\hat{G}_{x_1x_2}(f)e^{j2\pi f\tau}dt. \qquad \text{식(5)}$$

$$\psi_g(f) = 1/\left|G_{x_1x_2}(f)\right|. \qquad \text{식(6)}$$

이상에서 설명한 시간지연 추정방법의 성능에 영향을 미치는 요인들에 알아보도록 하겠다.

3. 공간의 잔향 특성에 따른 영향

3.1 *R/D* 비

R/D 비(reverberant-to-direct sound energy ratio, 잔향-직접 음 에너지 비)는 직접 음(direct sound) 에너지에 대한 잔향 음(reverberant sound) 에너지의 비로 정의되며 잔향 강도를 나타낸다.[9] *R/D* 비는 정규화된 실내 충격 응답 함수, $h(t)$ 을 이용하여 다음과 같이 나타낼 수 있다.

$$R/D\ ratio = 10\log\frac{\int_{0^+}^{\infty} h^2(t)dt}{\int_{0^-}^{0^+} h^2(t)dt}dB. \qquad \text{식(7)}$$

여기서 분모 항은 직접 음 에너지를 의미하고, 분자 항은 잔향 음 에너지를 의미한다. 본 연구에서는 측정된 신호의 시간 구간에 따른 *R/D* 비와 시간지연 추정 성능의 관계를 알아보기 위해 직접 측정된 신호를 직접 음과 잔향 음 부분으로 나눠 이 둘의 에너지 비로부터 *R/D* 비를 구하고자 한다. 마이크로폰으로부터 측정된 신호에는 직접 음과 잔향 음이 모두 포함되어 있기 때문에, 이 둘을 직접적으로 분리할 수가 없다. 따라서 실제 공간에서 측정된 실내 충격 응답 함수를 직접 음 부분과 잔향 음 부분으로 분리하여 임의의 음성신호를 convolution 함으로써 측정될 위치에서의 직접 음과 잔향 음 신호를 생성해 정의에 따라 *R/D* 비를 구하고자 한다.

3.2 *R/D* 비 계산을 위한 실내 충격 응답 함수 분리

실내 충격 응답 함수는 $4m \times 5m \times 3m$ 크기의 측정 공간에서 그림 1 과 같은 장비들을 이용하여 측정하였다. signal generator 에서 20H~20kH random white noise 를 speaker 를 통해 내보내주고, 이를 두 개의 마이크로폰을 이용하여 그 신호를 측정한 후, 입력신호와 측정된 신호 사이의

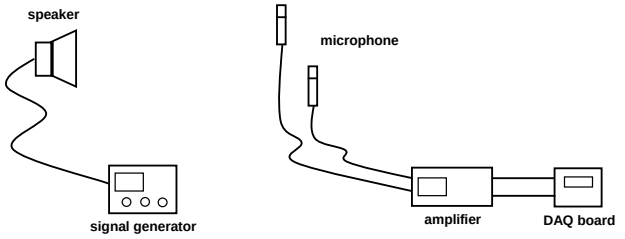
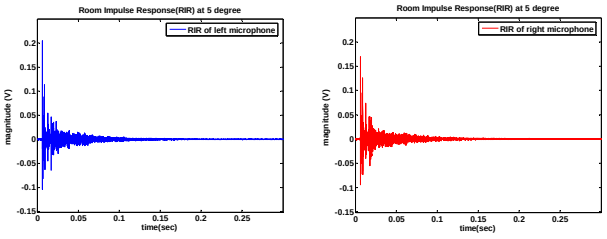


Fig.1 Equipments to measure room impulse response

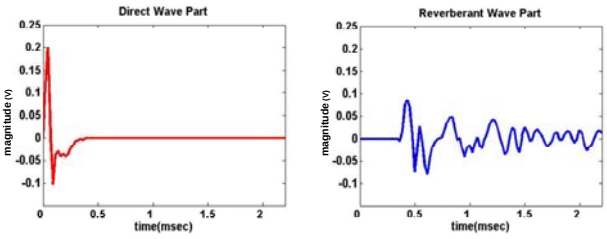
전달함수를 구하고, 측정된 전달함수의 역 퓨리에 변환을 통해서 실내 충격 응답 함수를 계산하였다. 시간지연 추정을 위해 두 개의 마이크로폰을 이용하였으며, 시간지연이 발생되도록 스피커의 위치는 한 쪽 마이크로폰으로 5° 치우치게 위치시켰다. 위와 같은 실험으로 측정된 두 개의 실내 충격 응답 함수는 그림 2 와 같다. 실제 측정된 실내 충격 응답 함수는 1.5 초이나, ISO 규정에 따라 계산된 잔향 시간이 약 0.28 초이므로, 그림 2 에서는 0.3 초까지만 나타내었다.

측정된 실내 충격 응답 함수를 직접 음 부분과 잔향 음 부분으로 나누기 위해서는 식(7)의 정의와 같이 실내 충격 응답 함수를 $0^- \sim 0^+$ 와 $0^+ \sim \infty$ 시간 부분으로 나눠야 한다. 하지만 이는 물리적으로 분리가 불가능 하기 때문에, 그림 3 과 같이 직접 음 부분은 첫 번째 peak 와 notch 이후에 실내 충격 응답 함수 값이 영(zero)이 될 때까지로 정하였고, 잔향 음 부분은 직접 음 부분



(a) (b)

Fig.2 Room impulse response; (a) left microphone, (b) right microphone



(a) (b)

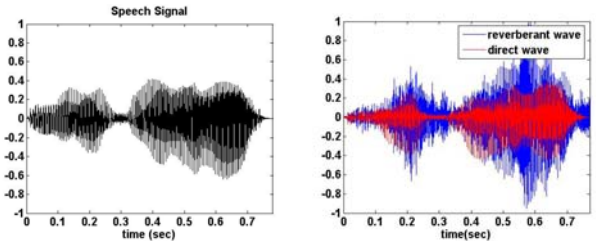
Fig.3 Separation of room impulse response; (a) direct wave (b) reverberant wave

이후부터 나머지 실내 충격 응답 함수 부분으로 하였다. 이는 실내 충격 응답 함수 측정에 사용된 입력 신호인 random white noise 가 모든 주파수 대역을 포함하고 있다면, 이론상 실내 충격 응답 함수의 직접 음 부분은 Dirac delta function 의 형태를 띠게 되지만, 실험에서는 제한된 주파수 대역의 random white noise 신호를 특정 샘플링 주파수로 측정하기 때문에 이를 이용할 경우 sinc function 의 형태로 나타나므로 이와 같이 분리하였다.

직접 음 부분과 잔향 음 부분으로 분리된 실내 충격 응답 함수에 임의의 음성신호(“안녕”)를 convolution 하여 마이크로폰 위치에서 측정된 직접 음과 잔향 음에 해당하는 신호들을 만들어 내면 그림 4 와 같다.

3.3 *R / D* 비와 시간지연 추정 성능

앞서 설명한 방법으로 만들어진 신호들의 직접 음과 잔향 음을 이용하여 일정한 시간간격(약 23msec, 512 samples)마다 *R/D* 비를 구해보면 그림 5 와 같다. 또한 동일한 시간간격(frame)의 두 신호들을 이용하여 상호상관값을 계산한다. 이 때 직접 음만을 이용할 경우와 잔향 음이 포함되어있는 전체 신호를 이용한 경우에 대해 각각 계산하면 그림 6 과 같다. 그림 6 에서 x 축은 각 시간 별로 상호상관값을 구하기 위한 시간간격, y 축



(a) (b)

Fig.4 Generation of direct and reverberant signal using room impulse response; (a) original speech signal, (b) generated signals

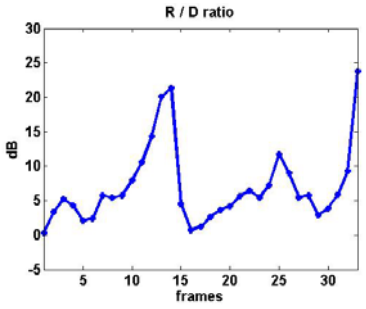
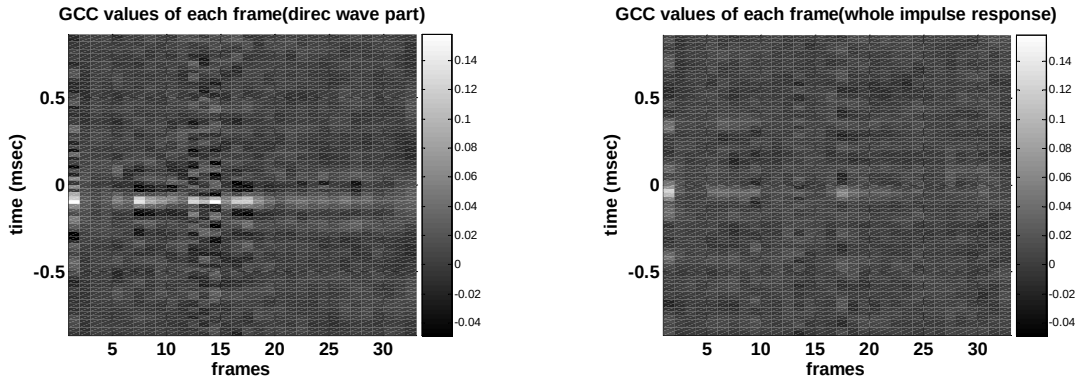
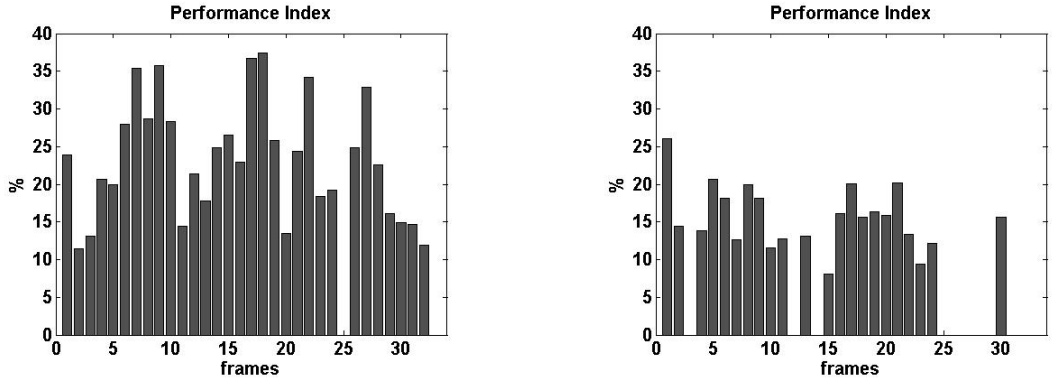


Fig.5 ***R/D*** ratio about speech signal; 1 frame is about 23msec (512 samples when sampling frequency is 22050H)



(a) (b)

Fig.6 Cross correlation of each frame about speech signal; (a) case of direct wave only, (b) case of whole wave (included reverberant wave)



(a) (b)

Fig.7 Performance index about speech signal; (a) case of direct wave only, (b) case of whole wave (included reverberant wave)

은 상호상관 값의 시간지연을 나타내며, 상호상관 값의 크기는 밝기의 차이로 표현되어 있다. 즉 밝은 색으로 표현된 곳은 상호상관값이 큰 곳이기 때문에 시간지연을 명확하게 나타낸다고 할 수 있다. 그림 6 의 (a)에서는 신호의 직접 음만을 사용했기 때문에 실제 두 신호의 시간지연을 나타내는 곳에서 상호상관값이 명확하게 나타나지만, (b)에서는 잔향의 영향으로 인해 전체적으로 그 명확성이 떨어지며, 10~15 frame 사이에서는 그 차이가 두드러진다. 이를 그림 5 의 *R/D* 비와 비교해 봤을 때, 그 부분에서 상대적으로 *R/D* 비가 크게 나타난다. 즉, *R/D* 비가 큰 부분에서는 시간지연 추정을 위한 상호상관값의 명확성이 떨어지며 그로 인해 추정오차가 발생하게 됨을 알 수 있다. 여기에서 시간지연 추정 성능을 정량적으로 나타내기 위해서 성능지표(performance index, PI)를 식(8)과 같이 정의한다.

$$\mathrm{PI}(\%) = \frac{\sum_{\tau_T-\varepsilon}^{\tau_T+\varepsilon} \left| Cross\ Correlation \right|}{\sum_{-\tau}^{\tau} \left| Cross\ Correlation \right|} \times 100 \quad \text{식(8)}$$

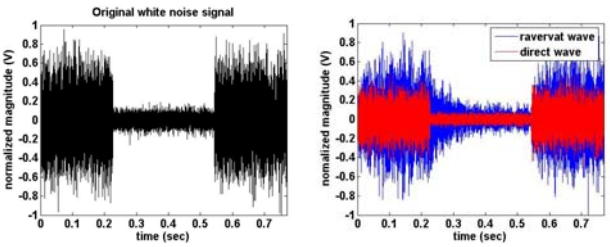
τ 는 마이크로폰 사이의 거리에 의해 물리적으로 발생될 수 있는 최대 시간지연이고, τ_T 는 음원의 위치에 따른 실제 시간지연이다. 또 ε 은 시간지연 값의 신뢰구간을 정의하는 변수로 본 연구에서는 τ 의 5%로 정의하였다. 위에서 정의한 성능지표는 측정된 신호에서 계산된 시간지연이 정의된 신뢰구간에 존재할 때만 값을 가지며, 그렇지 않은 경우에는 0 의 값을 가진다. 성능지표는 발생할 수 있는 상호상관값들의 전체 절대값 중에서 신뢰 구간의 절대값들이 차지하는 비율이며, 백분율로 나타난다. 여기에서 상호상관값들의 절대값으로 정의한 이유는 상호상관계수의 음의 값들도 그 만큼의 의미를 가지고 있기 때문이며, 성능지표는 측정되는 신호의 환경에 따라 기 기준이 변할 수 있는 상대적인 값이다. 그림 6 에서 구해진 각 frame 의 상호상관값들에 대한 성능지표를 구

해보면 그림 7 과 같다. 잔향 음까지 포함된 신호를 이용한 경우가 직접 음만을 이용한 경우보다 성능지표 관점에서 평균 11%의 성능저하를 보이고, 특히 *R/D* 비가 상대적으로 큰 10~15 frame 사이에서는 15%의 성능저하를 보이고 있다. 이상에서 *R/D* 비와 시간지연 추정성능 사에의 관계에 대해서 알아보았다. 하지만, 잔향의 영향이 포함되어있지 않는 직접 음만을 사용한 경우에도, 2~5, 28~33 frame 사이에서는 상호상관값의 명확성이 떨어지는데, 이는 다음 장에서 알아보도록 한다.

4. 신호의 특성에 따른 영향

앞장에서 잔향의 영향이 없는 직접 음만을 사용할 경우에도 상호상관값이 명확하게 나타나지 않는 부분이 있음을 알았다. 본 장에서는 이를 음성신호의 주파수 특성과 신호의 크기 변화의 관점에서 살펴보고자 한다.

시간지연 추정을 위해 사용된 음성신호의 스펙트로그램(spectrogram)을 구해보면 그림 8 과 같다. 그림 8 에서와 같이 음성신호의 주파수 성분이 넓게 분포하고 있는 6~20 frame 사이에서는 그림 6 의 (a)에서 보는 것과 같이 상호상관값의 최대값이 명확하게 나타나지만, 2~5 frame 사이에서와 같이 그렇지 않은 부분에서는 직접 음만을 이용한 경우에도 상호상관값의 명확성이 떨어짐을 알 수 있다. 이로부터 PHAT 가중치 함수를 적용한 일반화된 상호상관법을 이용하여 시간지연을 추정할 때, 넓은 범위의 주파수 성분을 갖는 음성신호를 이용할 경우 그 성능이 좋아짐을 알 수 있다. 이를 확인하기 위해 모든 주파수 대역을 가지는 random white noise 신호를 이용하여 앞의 과정을 반복해 보았다. 이 때, 신호의 크기 변화에 따른 시간지연 추정 성능도 알아보기 위해 그림 9 와 같이 신호의 크기가 변화는 신호를 이용하였다. 앞의 음성신호를 이용한 경우와 마찬가지로 분리된 실내 충격 응답 함수를 이용해 만들어진 신호로부터 각 시간구간별 *R/D* 비를 구하면 그림 10 과 같다. 그리고 동일한 간격의 두 신호를 이용하여 상호상관값을 계산하면, 직접 음만을 이용했을 경우에는 그림 11 의 (a)와 같고, 잔향음이 포함되어 있는 전체 신호를 이용할 경우에는 그림 11 의 (b)와 같이 나타난다. 여기에서 알 수 있듯이 전 주파수 대역을 포함하는 random white noise 신호의 경우에는 신호의 크기에 상관없이 상호상관값의 명확성이 뚜렷하게 나타남을 알 수 있다. 하지만 잔향의 영향이 포함되어 있는 경우에는 신호의 크기가 크다가 작아지는 부분인 10~13 frame 에서 명확성이 떨어지며, 신호의 크기가 작다가 커지게 되는 24 frame 에서는 명확성이 뚜렷하게 나타나게 된다. 이는 신호의 크기 변화에 따른 잔향의 영향이 달라져 *R/D* 비가 변화하기 때문이다. 이를 정량적으로 분석하기 위해서 앞서 정의한 성능지표를 각 frame 마다 계산해보면, 그림 12 와 같다. 직접 음만을 이용한 경우에는 신호의 주파수 특성이 고르게 분포하고 있기 때문에 모든 frame 에서 성능지표가 높게 나



(a) (b)

Fig.9 Generation of direct and reverberant signal using room impulse response; (a) original random white noise signal, (b) generated signals

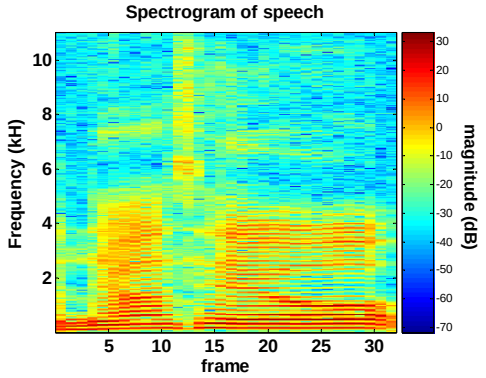


Fig.8 Spectrogram of speech signal

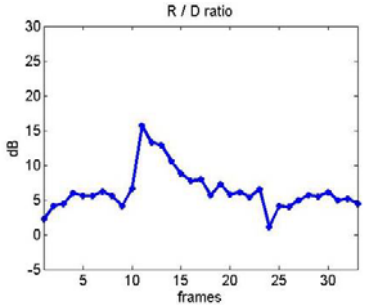
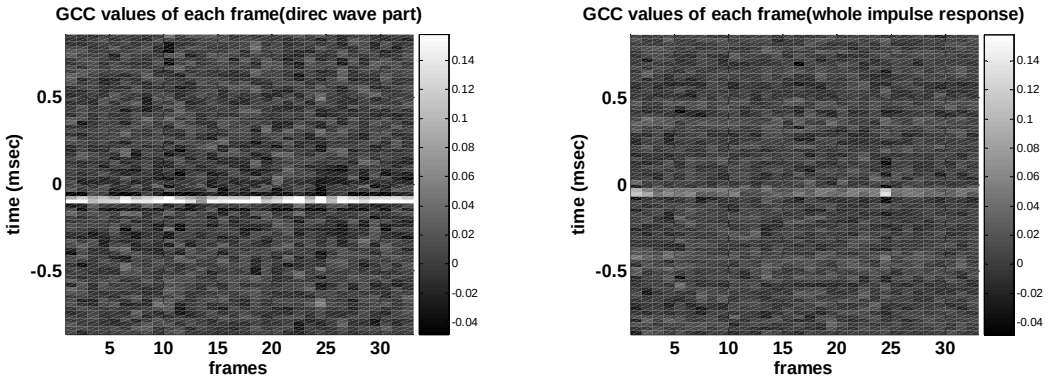
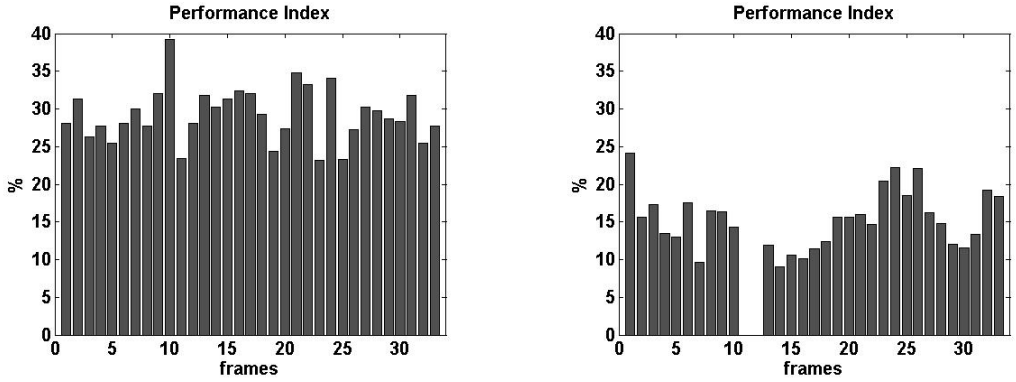


Fig.10 ***R/D*** ratio about random white noise signal; 1 frame is about 23msec (512 samples when sampling frequency is 22050H)



(a) (b)

Fig.11 Cross correlation of each frame about random white noise; (a) case of direct wave only, (b) case of whole wave (included reverberant wave)



(a) (b)

Fig.12 Performance index about white noise signal; (a) case of direct wave only, (b) case of whole wave (included reverberant wave)

타나지만, 잔향이 포함된 경우에는 성능지표 관점에서 평균 15% 성능저하가 발생함을 알 수 있다. 특히 신호의 크기 변화로 인해 *R/D* 비가 커지게 되는 10~13 frame 에서는 평균 25% 정도의 성능저하를 보이고, 신호의 크기가 작다가 커짐으로써 *R/D* 비가 작아지는 24 frame 에서는 12%의 성능저하만을 유발한다. 즉, 신호의 절대적인 크기는 시간 지연 추정 성능에 영향을 미치지 않지만, 시간에 따른 크기 변화는 결과적으로 잔향 특성에 영향을 미치므로 시간지연 추정 성능에 영향을 미치고 있음을 알 수 있다.

5. 결 론

본 연구에서는 PHAT 가중치 함수를 적용한 일반화된 상호상관법을 이용해 시간지연을 추정할 때, 그 성능에 미치는 영향들에 대해 알아보았다. 시간지연 추정성능은 상호상관값으로부터 정의되는 성능지표를 이용하여 정량적으로 나타내었다. 동일한 잔향시간을 갖는 공간에서 측정된 신호를 이용할 경우에 신호의 각 부분에 따라 시간지연 추정 성능이 달라지는데, 이를 직접 음 에너지에 대한 잔향 음 에너지의 비로 정의되는 *R/D* 비를 이용하여 분석해본 결과, *R/D* 비가 큰 부분에서는 시간지연 추정 성능이 나빠짐을 알 수 있었다. 또한 측정된 음성신호가 넓은 주파수 대역을 가질수록 좋은 성능을 보이며, 신호가 작다가 커지는 부분에서 상대적으로 *R/D* 비가 감소하여 시간지연 추정 성능이 좋아짐을 알 수 있었다.

후 기

이 논문은 2007 년도 정부(과학기술부)의 재원으로 한국과학재단의 국가지정연구실사업(R0A-2005-000-10112-0), 두뇌 한국 21 프로젝트 일환으로 수행하였음.

참고문헌

(1) Brandstein, M.S. and Silverman, H. F., 1997, “A practical methodology for speech source localization with

microphone arrays”, Computer Speech and Language, 11(2):91-126

(2) 권병호, 박영진, 박윤식, 2006, “로봇 시스템에 적용될 음원 위치 추정 방법”, 추계학술대회논문집, 한국소음진동공학회

(3) Wang, H. and Chu, P., 1997, “VOICE SOURCE LOCALIZATION FOR AUTOMATIC CAMERA POINTING SYSTEM IN VIDEOCONFERENCING”, Acoustics, Speech, and Signal Processing, ICASSP-97, Vol.1, pp. 187-190

(4) Knapp, C. H. and Carter, G. C., 1976, “The generalized correlation method for estimation of time delay”, IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. AS-24, No. 4, pp. 320-327

(5) Vaccaro, R. J., Ramalingam, C. S. and Tufts, D. W., 1992, “Least-squares time-delay estimation for transient signals in a multipath environment”, Journal of Acoustic Society of America, Vol. 92, No. 1, pp. 210-218

(6) Cabot, R. C., 1981, “A note on the application of the Hilbert Transform to time delay estimation”, IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-29, No. 3, pp. 607-609

(7) Chan, Y. T., So, H. C., and Ching, P. C., 1999, “Approximation maximum likelihood delay estimation via orthogonal Wavelet Trnasform”, IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-47, No. 4, pp. 1193-1198

(8) Brandstein, M. S., and Silverman, H. F., 1997, “A robust method for speech signal time delay estimation in reverberant rooms”, Acoustics, Speech and Signal Processing, Vol. 1, pp. 375-378

(9) 안상태, 1999 “자연스럽게 들리는 인공 잔향을 위한 FIR 필터 설계에 관한 연구”, 석사학위논문, KAIST