말소리와 음성과학 제 1 권 제 3 호 (2009)
pp. 95~??

# 확률적 목표 음성 검출을 통한 다채널 입력 기반 음성개선

## Probabilistic Target Speech Detection and Its Application to Multi-Input-Based Speech Enhancement

이 영 재[1] • 김 수 환[1] • 한 승 호[2] • 한 민 수[2] • 김 영 일[1] • 정 상 배[1]
Lee, Youngjae • Kim, Suhwan • Han, Seungho • Hahn, Minsoo • Kim, Young-Il • Jeong, Sangbae

**ABSTRACT**

In this paper, an efficient target speech detection algorithm is proposed for the performance improvement of multi-input speech enhancement. Using the normalized cross correlation value between two selected channels, the proposed algorithm estimates the probabilistic distribution function of the value from the pure noise interval. Then, log-likelihoods are calculated with the function and the normalized cross correlation value to detect the target speech interval precisely. The detection results are applied to the generalized sidelobe canceller-based algorithm. Experimental results show that the proposed algorithm significantly improves the speech recognition performance and the signal-to-noise ratios.

**Keywords: speech enhancement, speech recognition, microphone array beamforming**

## 1. 서 론

잡음제거의 주 목적은 음성기반 인터페이스의 성능을 높이는데 있다. 따라서, 핸즈프리키트(hands-free kits), 휴대 전화기, 보청기 같은 음성기반 응용 제품들에서 잡음제거 알고리즘이 필수적으로 사용되고 있다. 음성 신호는 가산성 혹은 합성곱(convolutive) 형태의 잡음에 의해서 왜곡이 된다. 그 중에서 잡음제거의 연구는 가산성 잡음의 제거에 주로 초점이 맞추어져 왔다. 그 이유는 음성기반 인터페이스 시장이 휴대형 전화기, PDA(personal digital assistants), 차량형 네비게이션 시스템에 집중되어 있기 때문이다.

잡음제거 기법은 가산성 잡음의 특징에 따라서 크게 단일채널 기반 접근법과 다채널 기반 접근법을 나눌 수 있다. 단일채널기반 잡음제거 알고리즘에서는 잡음의 통계량이 약 2~3 초 동안은 거의 변하지 않는다는 가정을 한다. 그런 후에 실기간으로 동작할 수 있는 VAD(voice activity detection) 알고리즘이 수행되어 잡음만이 존재하는 구간을 찾고 잡음의 통계치가 추정된다. 최종적으로 위너(Wiener) 혹은 칼만(Kalman) 필터에 의해서 잡음제거가 수행된다[1][2]. 이러한 방식은 차량 주행 잡음 혹은 컴퓨터 팬 잡음 같은 정상성 잡음의 제거에만 적합할 수 있다. 그러나 음성을 왜곡시키는 잡음의 통계량이 사람의 음성 혹은 음악소리 같은 비정상성 잡음일 경우에는 단일채널 기반 방식은 효율적으로 잡음제거를 수행할 수 없다. 그 이유는 비정상성 잡음 환경에서는 VAD 알고리즘이 믿을만한 성능을 보일 수 없기 때문이다. 다시 말하면, VAD 의 동작을 위해서 그 입력 값으로 에너지 궤적 및 주파수 특징 등을 사용하게 되지만 비정상성 잡음환경에서는 음성의 에너지 궤적과 잡음의 에너지 궤적간의 차이를 구분하기 힘들기 때문이다. 설령, VAD 가 음성과 잡음 구간을 잘

1) 경상대학교 전자공학과(공학연구원) jeongsb@gnu.ac.kr, 교신저자
2) 한국과학기술원 정보통신공학과

접수일자: 2009 년 8 월 18 일
수정일자: 2009 년 9 월 11 일
게재결정: 2009 년 9 월 14 일

분간할 수 있다 하더라도 잡음의 통계치는 심지어 음성 구간에서도 변할 수 있으므로 단일 채널 기반 잡음제거 방식은 최적의 성능을 기대하기 어렵다. 단일채널 방식의 여러 문제점을 극복하기 위해서 빔포밍(beamforming)과 BSS(blind source separation) 등의 다채널 입력 기반의 잡음제거 기법이 연구되어 왔다[4]-[7]. 일반적으로 빔포밍 알고리즘은 실제 환경에서 BSS 에 비해서 더 좋은 성능을 보인다고 알려져 있는데 빔포밍 기법은 목표 음원의 위치 정보를 부가적으로 이용하기 때문이다. 여러 가지 빔포밍 알고리즘 중에서 GSC(generalized sidelobe canceller) 기반의 방식이 가장 널리 쓰이고 있다. GSC 기반 잡음 제거에서 성능에

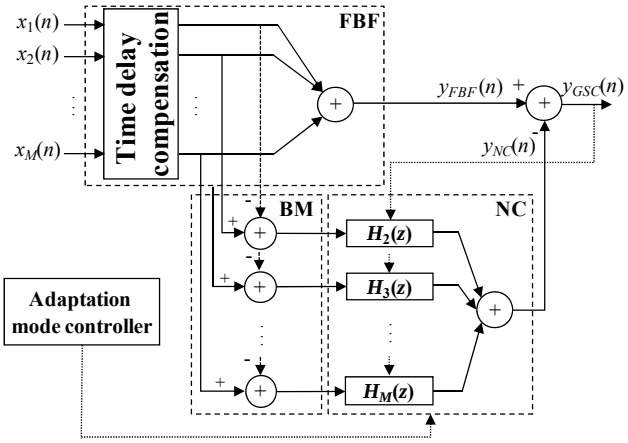


그림 1. GSC 기반 잡음제거 알고리즘의 일반적인 구조
Figure 1. Overall structure of the GSC-based speech enhancement algorithm

큰 영향을 미치는 모듈 중의 하나가 목표 음성 검출 알고리즘이다. 본 논문에서는 정밀한 목표 음성 검출 성능을 얻기 위해서 LPC(linear predictive coding) 잔차 신호에서 계산한 NCC(normalized inter-channel cross correlation)를 이용한다. 이로부터 잡음구간의 NCC 확률분포를 추정하고 이를 이용하여 전체 구간의 LL(log-likelihood)을 구한다. 최종적으로 LL 값이 임계치 이하이면 음성 구간으로, 이상이면 잡음 구간으로 판정하여 GSC 내부에 있는 적응 필터의 동작 모드를 결정한다.

본 논문의 구성은 다음과 같다. 제 2 장에서 기존의 연구를 소개하고, 제 3 장에서 제안된 알고리즘에 대해서 설명한다. 제 4 장에서 실험 결과에 대해서 논하고 마지막으로 제 5 장에서 본 논문의 결론을 맺는다.

## 2. 관련 연구

그림 1 에서 GSC 기반 잡음제거 알고리즘의 일반적인 구조를 나타내었다. GSC 알고리즘은 크게 FBF(fixed beamforming) 모듈, BM(blocking matrix) 모듈, NC(noise cancellation) 모듈 등의 3 부분으로 나눌 수 있다. 그림 1 에서 $x_i(n)$은 $i$번째 마이크로폰에서 수신된 신호이며, $M$ 은 마이크로폰의 총 개수이다. $y_{FBF}(n)$은 FBF모듈의 출력, $H_i(z)$는 NC 모듈에 있는 $i$ 번째 적응필터, $y_{NC}(n)$은 NC 모듈의 출력, $y_{GSC}(n)$은 GSC에 의한 최종 잡음제거 결과이다. GSC에 의한 잡음제거를 위해서 선행적으로 주어지는 목표 음원의 DOA(direction of arriaval)를 이용하여 목표 신호의 입력 채널간 시간차를 보정해주는 TDC(time delay compensation) 를 수행한다. 그 후, 모든 채널의 샘플 값을 더해서 FBF의 출력을 구한다. 기본적으로 FBF모듈은 목표 신호의 에너지를 크게 하는 역할을 하는데 목표 음원의 위치 이외에서 들어오는 신호가 잔존하는 것을 피할 수는 없다. 여기서, FBF 결과에 섞여있는 잔존 잡음은 BM 모듈에서 추정한 잡음 참조신호를 이용하여 제거가 가능하다. 즉, BM 모듈의 잡음 참조 신호는 NC 모듈의 적응 필터에 의해서 $y_{FBF}(n)$에 섞여 있는 잡음과 유사한 형태로 변형된다. 그런데, NC 모듈의 적응 필터의 계수 갱신이 목표 음성구간에서도 일어나게 된다면 잡음 제거 후의 음성 신호에 심각한 왜곡이 발생될 수 있다. 그 이유는 실제 상황에서 두 가지로 생각할 수 있다. 첫 번째 이유는 BM 모듈에서 추정한 잡음 참조 신호에는 순수 잡음만이 모델링되지 않기 때문이다. 즉, DOA 정보를 이용하여 FBF 모듈에서 채널간 목표 음성 신호에 대한 시간차를 보정했음에도 목표 음원과 마이크로폰간의 경로차 혹은 ADC(analog-to-digital converter)를 포함한 마이크로폰 특성 차이 때문에 단순 뺄셈 만으로는 BM 모듈에서 완벽한 목표 음성신호의 소거가 일어날 수 없다. 두 번째 이유는 참조 잡음 신호가 $y_{FBF}(n)$에 존재하는 목표 음성 신호와 통계적 상호상관도가 어느 정도 존재할 수 있기 때문이다. 잡음과 목표 음성간의 상관도는 잡음이 인간의 음성 및 음악 신호일 때 주파수 영역에서의 분포가 비슷하여 특히 크게 나타날 수 있다. 따라서, GSC가 실제 환경에서 더 좋은 성능을 갖게 하기 위해서 AMC(adaptation mode controller)를 부가적으로 사용하게 된다. AMC는 현재 처리해야 할 구간이  목표 음성 구간인지 잡음 구간인지를 판별하여 NC 모듈의 동작 모드를 결정한다. 만약, 잡음 구간일 경우에 NC 모듈의 적응 필터의 계수 갱신을 실시하고 목표 음성 구간일 경우에는 갱신을 수행하지 않는다. 만약, NC 모듈의 적응 필터 계수가 갱신되지 않으면 잡음 제거의 양은 줄어들게 된다. 그러나, 목표 음성 신호의 왜곡을 방지할 수 있으므로 실보다는 득이 크다고 볼 수 있다. 이러한 GSC의 전체적인 동작을 봤을 때, 잡음 제거 후의 신호를 청취하면 목표 음성 구간에서 갑자기 잡음의 소리가 커지게 될 것임을 예상할 수 있다.

여러 가지 방식으로 AMC의 설계 기법이 연구되어 왔는데, 보편적으로 널리 쓰이는 방식으로서 $y_{FBF}(n)$의 에너지 궤적을 이용하는 방법과 $y_{FBF}(n)$과 $y_{NC}(n)$의 통계적 상호상관도를 이용하는 방법을 2.1, 2.2 에 소개하였다.

### 2.1 고정 빔포밍 후의 에너지 궤적을 이용하는 방법

그림 1 에서 고정 빔포밍을 수행하면 이상적인 상황에서 목표 음성 신호가 $M$배만큼 커질 수 있다. 반면 잡음은 채널 별로 신호의 위상이 차이가 나므로 $M$ 배보다는 작은 증폭 효과를 얻는다. 즉, $y_{FBF}(n)$은 $x_i(n)$ 보다 높은 SNR(signal-to-noise ratio)를 갖게 된다. 따라서, 단순 에너지의 크기로만 목표 음성 구간과 잡음 구간을 분류하기를 원한다면 $y_{FBF}(n)$ 신호에서 에너지 궤적을 구해서 분석하는 것이 더 높은 성공률을 기대할 수 있다. $y_{FBF}(n)$에서 단위 분석 프레임마다 구해진 에너지 값으로부터 단순 임계치를 이용하여 주어진 임계치 이상이면 목표 음성구간, 그렇지 않으면 잡음 구간으로 판정할 수 있다. 만약, 더 높은 성공률을 얻으려면, 음성 부호화 알고리즘 혹은 음성 인식 알고리즘에서와 같이 인간의 음성 특성을 이용할 수 있다.

### 2.2 NC 동작 전후의 통계적 상호상관도를 이용하는 방법

NC 모듈에서 추정된 잡음이 $y_{FBF}(n)$에서 제거될 때에 $y_{FBF}(n)$의 통계적 변화는 목표 음성구간에 비해서 잡음 구간이 더 크다고 볼 수 있다. 즉, $y_{FBF}(n)$과 잡음 제거 후의 신호인 $y_{GSC}(n)$의 상호 상관도는 잡음 구간에서 작고 목표 음성구간에서 크다. 이러한 정보를 이용하여 목표 음성구간과 잡음 구간을 나눌 수 있다. 정규화된 상호 상관도 $\rho_{FBF,GSC}$ 의 정의를 수식 (1)에 나타내었다[6].

$$\rho_{FBF,GSC} = \frac{P_{FBF,GSC}(n)}{\sqrt{P_{FBF}(n)P_{GSC}(n)}} \tag{1}$$

$P_{FBF,GSC}(n)$, $P_{FBF}(n)$, $P_{GSC}(n)$ 은 상호 상관도, $y_{FBF}(n)$의 전력, $y_{FBF}(n)$의 전력을 나타낸다. 수식 (2)-(4)에 자세히 나타내었다.

$$P_{FBF,GSC}(n) = (1-\lambda)P_{FBF,GSC}(n-1) + \lambda y_{FBF}(n)y_{GSC}(n) \tag{2}$$

$$P_{FBF}(n) = (1-\lambda)P_{FBF}(n-1) + \lambda y_{FBF}^2(n) \tag{3}$$

$$P_{GSC}(n) = (1-\lambda)P_{GSC}(n-1) + \lambda y_{GSC}^2(n) \tag{4}$$

여기서 $\lambda$ 는 1 보다 작은 상수이다. 따라서, $\rho_{FBF,GSC}$ 는 목표 음성 구간에서는 1 에 가까운 값을 갖고, 잡음 구간에서는 0 에 가까운 값을 갖게 됨을 예상할 수 있다. 이때의 AMC 는 $\rho_{FBF,GSC}$ 가 주어진 임계값을 넘을 경우에 NC 모듈 적응 필터의 계수 갱신을 수행하지 않게끔 한다.

## 3. 제안된 AMC 알고리즘

기존의 목표 구간 검출 알고리즘 중에서 2.1 절의 $y_{FBF}(n)$의 에너지를 이용하는 방식은 입력 SNR이 낮아질수록 성능의 열화가 심해지는 단점을 예상할 수 있다. 비정상성 잡음환경에서 입력 SNR이 낮을 경우에 $y_{FBF}(n)$의 에너지 궤적은 배경 잡음

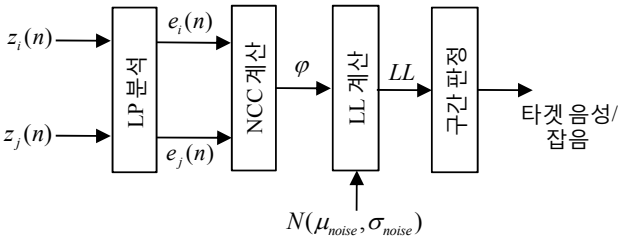


그림 2. 제안된 AMC 알고리즘의 구조
Figure 2. Structure of the proposed AMC algorithm

구간과 목표 음성구간을 분류하기 어려울 정도로 변동이 심할 수 있다. 또한, 실제 환경에는 목표 음원과 마이크로폰 배열 사이에 존재하는 경로차 때문에 $y_{FBF}(n)$의 목표 음성 신호가 사용되는 마이크로폰의 개수의 배수보다 상대적으로 작은 증폭비를 갖는다. 2.2 절에서 설명한 NC 동작 전후의 통계적 상관도를 이용하는 알고리즘은 입력 신호의 방향 정보가 통계적 상관도에 어느 정도 반영될 수 있으므로SNR이 낮더라도 2.1 절의 방식보다 좋은 성능을 가질 수 있다. 일반적으로 빔포밍 기반의 음성 개선 방식은 반향이 많이 존재하는 원거리 음성 입력에 대한 동작을 고려하는데 2.1 절, 2.2 절에서 설명한 기존의 목표 음성 검출 알고리즘은 반향에 대한 고려를 하지 않고 있다. 따라서, 본 연구에서는 목표 음성 구간의 판정에 있어서 방향 정보를 이용하기 위해서 채널간 상호 상관도를 이용하는 기법을 기본으로한다. 여기에 반향이 존재하는 조건에서 좀 더 정확한 검출을 위해서 입력 신호의 LP(linear prediction) 잔차 신호에서 상호 상관도를 추정하도록 한다. 그림 2 에서 제안된 방식의 AMC 구조를 나타내었다. 그림 2 에서 $z_i(n)$, $z_j(n)$ 는 $i$ 번째와 $j$ 번째 채널 입력 신호에서 TDC가 수행된 결과이다. $e_i(n)$, $e_j(n)$ 는 각각의 LP 잔차신호이며 $\varphi$ 는 단구간에서 계산된 NCC(normalized cross correlation) 값이다. LL(log-likelihood)는 $\varphi$ 를 잡음 NCC의 정규분

포인 $N(\mu_{noise},\sigma_{noise})$ 에 대입해서 구해진다. 제안된 알고리즘은 기본적으로 채널간 NCC를 측정하여 입력 구간의 판정을 수행하게 된다. 즉, TDC가 수행된 후에는 각 채널에서 목표 음성 신호의 시간차가 존재하지 않으므로 목표 음성 구간에서는 NCC가 크게 측정되고 잡음 구간에서는 상호 상관도가 낮게 측정된다. 3.1-3.4 절에 세부적인 동작을 나타내었다.

### 3.1 LP 분석

LP 분석을 위해서 먼저 입력 신호에 Hamming 창을 곱하고 수식 (5)를 이용하여 자기 상관 계수(autocorrelation coefficient)를 추정한다.

$$r_i(k) = \sum_{n=k}^{N-1} z_i'(n) z_i'(n-k) \tag{5}$$

여기서, $r_i(k)$ 는 $i$ 번째 채널에 대한 $k$ 번째 차수의 자기 상관 계수, $z_i'(n)$ 는 $z_i(n)$ 에 Hamming 창이 곱해진 신호이다. $N$ 은 단구간 분석 프레임의 데이터 수 이다. 수식 (5)를 통해서 구해지는 $r_i(k)$ 는 Levinson-Durbin 알고리즘에 의해서 LP 계수로 변환된다 [8]. 추정된 LP 계수가 $a_i(k)$ 일 때, LP 잔차 신호는 수식 (6)을 통해서 얻어진다.

$$e_i(n) = z_i(n) - \sum_{k=1}^{P} a_i(k) z_i(n-k) \tag{6}$$

여기서 $P$ 는 LP 계수 추출 차수이다. 일반적으로 반향 환경에서의 음향신호는 파형 자체가 넓게 퍼지게 되어 채널간 시간차가 잘 나타나지 않게 된다. 이러한 문제는 빔포밍 알고리즘으로 하여금 목표 음성 신호와 잡음 신호의 구분을 어렵게 하는 주요 원인이다. 만약, 입력 신호에서 LP 잔차 정보를 이용하게 되면 신호가 다소 뾰족하게 바뀌므로 반향 성분의 악영향을 어느 정도 완화시킬 수 있다. 이는 NCC 궤적의 측면에서 목표 음성 구간을 더욱 명확히 나타낼 수 있도록 한다.

### 3.2 NCC 계산

수식 (7)에 선택된 두 개의 채널 입력이 주어졌을 때의 NCC 계산 방법을 나타내었다.

$$c_t = \frac{\sum_{n=0}^{L-1} e_{i,t}(n) e_{j,t}(n)}{\sqrt{\sum_{n=0}^{L-1} e_{i,t}^2(n)} \sqrt{\sum_{n=0}^{L-1} e_{j,t}^2(n)}} \tag{7}$$

여기서, $c_t$ 는 $t$ 번째 단구간 입력 프레임에 대한 NCC를 나타낸다. $e_{i,t}(n)$, $e_{j,t}(n)$는 $i$번째 및 $j$ 번째 채널에 대한 LP 잔차 신호이며 $L$ 은 단구간 입력 프레임의 길이를 나타낸다.

### 3.3 LL 계산

LL 계산을 위하여 선행적으로 잡음 구간 NCC를 가우시안 분포로 모델링하고 분포의 평균치($\mu_{noise}$) 와 분산치($\sigma_{noise}$)를 순수 잡음 구간에서 추정한다. 본 연구에서는 초기 200 ms입력 구간에는 목표 음성 신호가 존재하지 않는 순수 잡음 구간으로 간주한다. 수식 (8)에 LL 계산식을 나타내었다.

$$\begin{aligned} LL_t &= \ln\left( \frac{1}{\sigma_{noise}\sqrt{2\pi}} \exp\left( -\frac{(c_t - \mu_{noise})^2}{2\sigma_{noise}^2} \right) \right) \\ &= G - \frac{(c_t - \mu_{noise})^2}{2\sigma_{noise}^2} \end{aligned} \tag{8}$$

여기서 $G = -\ln(\sigma_{noise}) - 0.5 \cdot \ln(2\pi)$ 이다.

### 3.4 구간 판정

임의의 단구간 입력이 목표 음성을 포함하고 있는지의 여부는 그 구간에서 구해진 LL 값이 주어진 임계치 이상인지를 체크하여 판정한다. LL 값에 대한 임계치는 초기 200 ms 입력에 대한 LL 값의 평균치에 특정 값을 더한 것으로 한다. 이 때, 초기 입력 구간은 3.3 절에서 잡음 구간 NCC의 분포 파라미터 추정이 이루어진 구간과 같다. 초기 입력 구간에 대한 LL 값에 대한 평균치 $LL_{mean}$은 수식 (9)와 같다.

$$\begin{aligned} LL_{mean} &= \frac{1}{T_{init}} \sum_{t=0}^{T_{init}-1} LL_t \\ &= \frac{1}{T_{init}} \sum_{t=0}^{T_{init}-1} \left( G - \frac{(c_t - \mu_{noise})^2}{2\sigma_{noise}^2} \right) \\ &= G - \frac{1}{2} \end{aligned} \tag{9}$$

여기서, $T_{init}$는 초기 200 ms 에 수신된 단구간 입력 프레임의 수이며 수식 간략화를 위해서 $\sum_{t=0}^{T_{init}-1} (c_t - \mu_{noise})^2 \Big/ T_{init} = \sigma_{noise}^2$ 임이 이용되었다. 최종적으로 구간 판정을 위한 임계치는 수식 (10)과 같이 설정한다.

$$LL_{TH} = LL_{mean} - \beta \tag{10}$$

$\beta$ 는 목표 음성 검출을 구하기 위하여 실험적으로 설정해야 하

는 상수이다. 3.3 절에서 구한 분포의 파라미터는 잡음 구간에 대한 것 들이므로 LL 값은 잡음 구간에서 높게 나오고 목표 음성 구간에서 낮게 나온다. 따라서, 수식 (10)을 이용할 때는 임의의 입력에 대한 LL 값이 $LL_{TH}$이하일 때에 목표 음성 구간으로 판정하고 그렇지 않을 경우에는 잡음 구간으로 판정한다.

## 4. 실험 및 결과

### 4.1 다채널 DB 수집 및 실험 조건

목표 음성과 잡음 신호는 6 m X 5 m 크기의 일반 가정의 거실에서 각각 따로 녹음되었다. 잡음이 섞인 목표 음성을 얻기

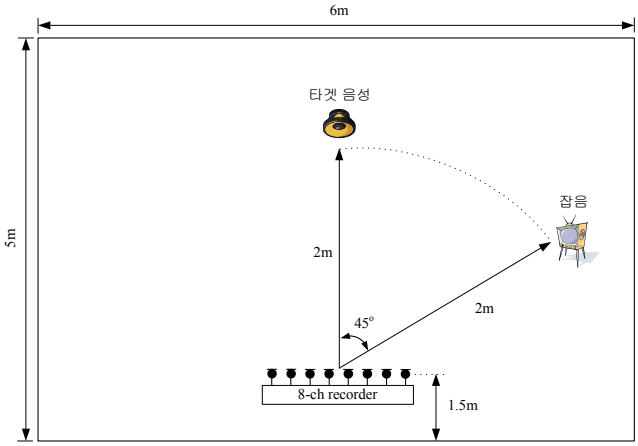


그림 3. 음원 및 녹음 장비의 전체적인 배치도
Figure 3. Overall placement of the sound sources and the recording device

위해서 잡음 신호의 구간은 무작위로 선택되었으며 원하는 입력 SNR 을 얻기 위해 크기가 조절된 후에 목표 음성 신호에 더해진다. 데이터의 녹음을 위하여 한국전자통신연구원에서 지능형로봇에 장착될 용도로 제작된 장비가 사용되었으며 마이크로폰의 수는 8 개이다. 마이크로폰은 선형으로 배치되었으며 마이크로폰 간의 간격은 2 cm 로 설정하였다. 목표 음성 신호로는 10 명의 남성과 10 명의 여성이 발성한 총 904 개의 한국어 PBW(phonetically balanced world)가 사용되었으며 고품질 음향 스피커를 통해서 재생되었다[10]. 잡음신호로 Kelly Clarkson 이 부른 'Because of you' 가 42 인치급 프로젝션 TV 에서 재생되었다. 모든 데이터는 표본화율 16 kHz, 해상도 16 bit 로 저장되었다. 음원 및 장비의 전체적인 배치는 그림 3 과 같고 지면에서 음원 및 녹음 장비와의 수직 높이는 약 1 m 였다.

### 4.2 성능 평가

제안된 알고리즘은 2.1 절에서 설명한 고정 빔포밍의 에너지 궤적을 이용하는 방법과 2.2 절에서 설명한 NC 전후의 상관도를 이용한 방법과 비교되었다. 고정 빔포밍의 에너지 궤적을 이용할 때는 그림 1 의 $y_{FBF}(n)$에서 10 ms 단위로 에너지를 추출하여 참고문헌 [1]에 있는 VADNest 알고리즘을 사용하여 목표 음성 구간 및 잡음 구간을 판정한 후 NC 모듈 내의 필터 적응 여부를 판단한다. NC 전후의 상관도를 이용할 때에는 수식 (1)-(4)가 그대로 적용되었으며 $\lambda$=0.01 로 설정하였다. 목표 음성 검출을 위한 $\rho_{FBF,GSC}$ 의 임계치는 실험적으로 최적화되었으며 본 연구에서는 0.55 가 사용되었다. NC 필터의 길이는 동일하게 127 이며, 잡음 구간일 경우에는 NLMS(normalized least-mean-square) 알고리즘에 의해서 계수 값이 갱신된다. 각 필터의 학습률은

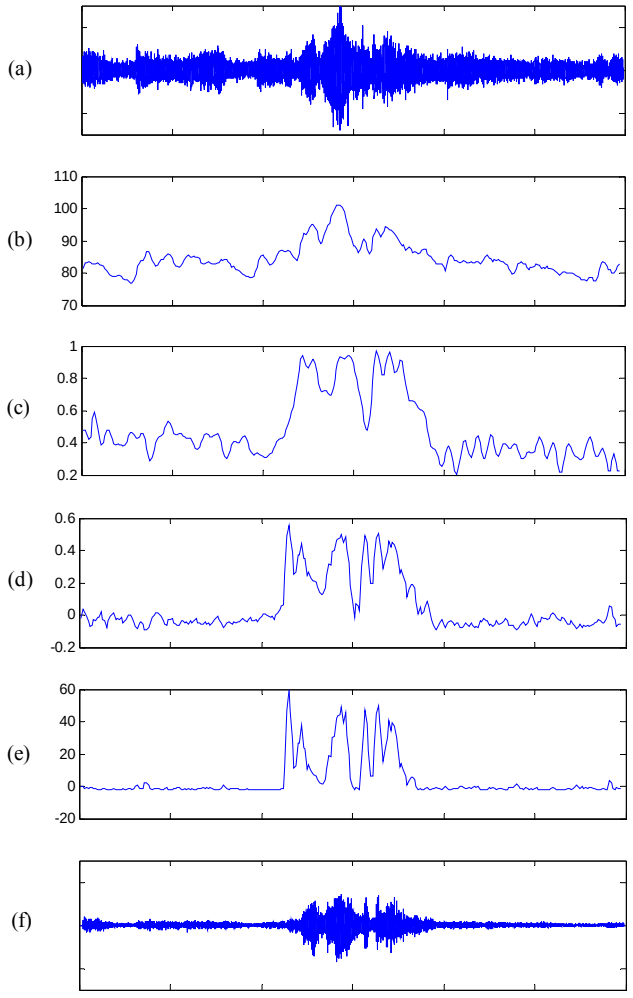


그림 4. (a) SNR 5dB 의 입력 신호, (b) 고정 빔포밍 결과의 로그 에너지 궤적, (c) NC 동작 전후의 상호 상관도 궤적, (d) 제안된 방식에 의한 NCC 궤적, (e) 제안된 NCC 의 LL 궤적, (f) 제안된 방식에 의한 잡음제거 결과, (*수평축의 한 칸은 1 초에 해당함)
Figure 4. (a) Input signal with 5dB SNR, (b) Log energy contour of fixed beamforming result (c) Cross-correlation contour between $y_{FBF}(n)$ and

$y_{GSC}(n)$, (d) NCC contour by the proposed method, (e) LL contour of NCC by the proposed method, (f) noise reduction result by the proposed, (* One second for a tick on the horizontal axis)

동일하게 0.1 로 설정되었다. 본 연구에서는 목표 음원이 마이크로폰 배열의 정면에 위치하므로 그림 1 의 TDC 과정은 수행하지 않았다. 제안된 방식을 실험할 때에 LP 분석 차수는 18 이었고 매 10 ms 당 30 ms 의 분석 프레임에서 추출되었다. NCC 및 LL 역시 매 10ms 단위로 추출되었으며 수식 (10)의 $\beta$ 에 10 을 할당하였다. 그림 4 에서 SNR 5dB 정도의 입력 음성, 고정 빔포밍 결과의 에너지 궤적, NC 동작 전후의 상관도, 제안된 방식인 LP 잔차 신호의 채널간 상관도(NCC), NCC 에 대한 LL, 제안된 방식에 의한 잡음제거 결과를 차례로 나타내었다. 여기서, NCC 의 LL 궤적은 분석의 편의를 위해서 부호를 반대로 하여 그렸다. 그림 4 를 보면 제안된 방식을 사용할 경우에 가장 용이하게

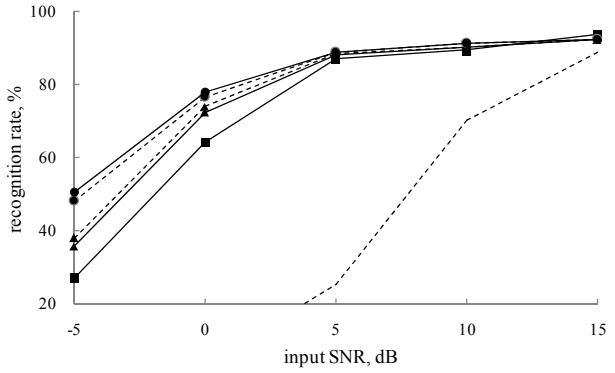


그림 5. 음성인식률 측정 결과(점선: 잡음 제거 전의 입력 신호, ■: 고정 빔포밍 결과의 에너지 궤적 이용, ▲: NC 동작 전후의 상호상관도 이용, ▲-점선: 입력 신호의 상호상관도와 LL 을 이용, ●-점선: LP 잔차신호의 상호상관도만을 이용, ●: 제안된 방식)

Figure 5. Speech recognition results(dashed: for input noisy signal, ■: for energy contour of the fixed beamforming result, ▲: for cross-correlation between yFBF(n) and yGSC(n), ▲-dashed: for the cross-correlation of input signals and the log-likelihood, ●-점선: for the cross-correlation of LP residual signals,●: proposed method)

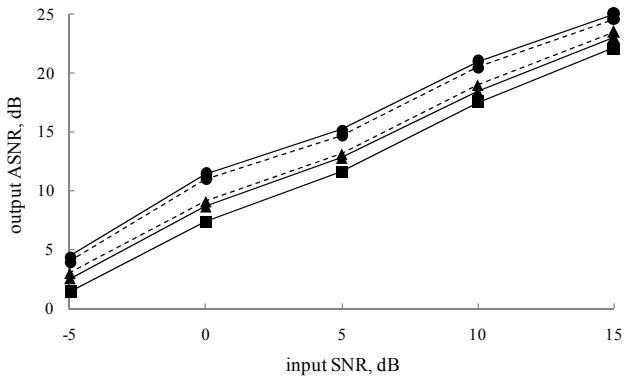


그림 6. 평균 SNR 측정 결과

Figure 6. Averaged SNR results

목표 음성 구간을 찾을 수 있 음을 알 수 있다. 특히, LP 잔차 신호의 NCC(그림 4(d))를 입력으로 하여 잡음 구간 NCC 의 분포를 이용한 로그 우도(그림 4(e))를 계산할 경우에 목표 음성 구간을 더욱 잘 검출할 수 있음을 알 수 있다. NC 동작 전후의 상호 상관도를 이용하는 방법은 고정 빔포밍 결과의 에너지 궤적을 이용하는 경우보다 더 좋은 결과를 얻을 수 있었지만 제안된 방식을 사용할 때와 비교해 보면 잡음 구간에서 궤적의 요동이 훨씬 심함을 알 수 있다. 제안한 알고리즘의 성능 평가를 위해서 음성인식률과 평균 SNR 이 사용되었다. 음성인식률 측정을 위해서 HMM(hidden Markov model) 기반의 음성인식기가 사용되었다[9]. Triphone 기반의 HMM 이 사용되었으며 음향 모델 훈련을 위한 DB 가 평균 20 dB 의 SNR 을 갖게 하기 위해서 음악

표 1. 전체 입력 SNR 에 대해서 평균 인식률

Table1. Average recognition rate for all input SNRs

| | |
|---|---|
| 잡음 제거 전의 입력 신호 | 37.61 % |
| 고정 빔포밍 결과의 에너지 궤적 이용 | 72.31 % |
| NC 동작 전후의 상호상관도 이용 | 75.73 % |
| 입력 신호의 상호상관도와 LL 을 이용 | 76.62 % |
| LP 잔차신호의 상호상관도만을 이용 | 79.34 % |
| 제안된 방식 | 80.22 % |

표 2. 전체 입력 SNR 에 대해서 평균 SNR 의 평균치

Table2. Average ASNR for all input SNRs

| | |
|---|---|
| 고정 빔포밍 결과의 에너지 궤적 이용 | 12.03 dB |
| NC 동작 전후의 상호상관도 이용 | 13.08 dB |
| 입력 신호의 상호상관도와 LL 을 이용 | 13.54 dB |
| LP 잔차신호의 상호상관도만을 이용 | 14.96 dB |
| 제안된 방식 | 15.39 dB |

잡음을 가산하였으며 가산된 잡음은 테스트 DB 구성을 위해 가산한 잡음과는 다르다. HMM 의 상태 수는 각각 3 개를 할당하였으며 상태의 분포는 공유된다. 본 실험에서 사용된 HMM 상태의 총 수는 1000 개였다. 음성인식의 특징 파라미터로서 MFCC (mel-frequency cepstral coefficient)와 단구간 에너지가 10 ms 단위로 추출되었으며 그것의 속도, 가속도 성분이 부가적으로 추출되어 총 39 차의 벡터가 사용되었다. 음성인식률 측정 시에 음성 구간은 수작업으로 검출되었다. 음성인식기는 HTK(HMM

training toolkit)으로 구현되었으며 인식대상 어휘 수는 452 였다 [11]. 평균 SNR 의 정의는 빔포밍에 의한 잡음 제거 후에 순수 잡음 구간의 평균 전력 대비 목표 음성 구간의 평균 전력으로 나타낼 수 있다. 수식 (11)에 자세히 나타내었다.

$$ASNR = \frac{P_s - P_v}{P_v} \tag{11}$$

$P_s = \frac{1}{N_s}\sum_{n\in\Phi_s} y_{GSC}^2(n)$ , $P_v = \frac{1}{N_v}\sum_{n\in\Phi_v} y_{GSC}^2(n)$ 이며 $\Phi_s$ , $\Phi_v$ 는 각각 목표 음성 및 잡음 구간을 나타낸다. 그림 5, 6 에 음성인식률과 평균 SNR 의 측정 결과를 나타내었다. 음성인식률 측정 결과 제안된 방식이 전체적으로 가장 좋은 성능을 보였다. 특히, SNR 이 낮아질수록 더 좋은 성능을 보임을 알 수 있었다. 입력 SNR 이 5 dB 일 때의 인식률은 잡음제거 전이 25. 1 %, 고정 빔포밍의 에너지 궤적을 이용한 방식이 87 %, NC 동작 전후의 상호 상관도 이용법이 87.9 %, 제안된 방식이 88.8 % 였다. 전체 입력 SNR 에 대해서 평균 인식률을 표 1 에 정리하였다. 그림 5 에서 LP 잔차신호 대신에 입력 신호의 상호상관도를 사용하고 3.3, 3.4 절의 확률적 검출 기법을 사용할 때와, LP 잔차 신호를 사용하고 확률적 검출 기법을 사용하지 않는 대신에 초기 200ms 구간의 순수 잡음구간에서 추정한 NCC 값의 평균 크기보다 두 배인 구간을 목표 음성구간으로 검출할 때의 성능을 부가적으로 나타내었다. 두 가지 요인이 성능 향상에 미치는 효과를 살펴보면 LP 잔차 신호를 사용하는 것이 로그 우도 기반의 확률적 검출 기법에 비해서 성능 향상에 더 큰 기여를 하고 있음을 확인할 수 있다. 평균 SNR 의 측정 결과도 음성인식률과 마찬가지로 제안된 방식이 가장 좋은 성능을 보였다. 각 입력 SNR 에서 제한된 방식은 NC 동작 전후의 상호 상관도를 이용할 때 보다 약 1.9 dB 의 향상을 얻을 수 있었으며 고정 빔포밍의 에너지 궤적을 이용하는 방법보다는 약 3.5 dB 의 향상을 얻을 수 있었다. 전체 입력 SNR 에 대한 평균 SNR 의 평균치를 표 2 에 정리하였다.

## 5. 결 론

본 연구에서는 다채널 마이크로폰 입력을 사용하여 GSC 기반의 빔포밍 알고리즘으로 잡음 제거를 수행할 때에 성능 향상을 위한 알고리즘을 제안하였다. 제안된 알고리즘에서는 목표 음성 신호의 채널간 상호 상관도를 더욱 명확히 얻기 위해서 LP 잔차 신호를 기본적으로 사용하였으며 배경 잡음의 분포 특성을 활용하기 위하여 확률 분포에 기반을 둔 로그 우도가 부가적으로 이용되었다. 기존 알고리즘과 음성인식률 및 평균 SNR 을 비교하였을 때 가장 좋은 성능을 보임을 알 수 있었다. 향후 연구로써 빔포밍 후에 위너 필터 혹은 칼만 필터를 이용하여 단일 채널 잡음 제거를 수행하여 더욱 향상된 성능을 얻을 수 있는지와 마이크로폰의 배치가 원형일 때의 성능을 측정하여 지능형 로봇에 장착이 가능한지를 타진해 볼 예정이다.

## 감사의 글

본 연구는 지식경제부 및 정보통신연구진흥원의 IT 성장동력 기술개발사업의 일환으로 수행하였음. [2008-F-037-01, u-로봇 HRI 솔루션 및 핵심소자 기술개발].

## 참 고 문 헌

[1] ETSI ES 202 212, (2005). Speech processing, transmission and quality aspects (STQ), v.1.1.2.

[2] Jeong, S. and Hahn, M. (2001). “Speech quality and recognition rate improvement in car noise environments”, Electronics Letters, Vol. 37, No. 12, pp. 801-802.

[3] Brandstein, M. and Ward, D. (2001). Microphone Arrays: signal processing techniques and applications, Springer-Verlag.

[4] Hoshuyama, O. et al. (1999). “A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters”, IEEE Trans. Signal Proc., Vol. 47, No. 10, pp. 2677-2688.

[5] S. Gannot et al. (2001). “Signal enhancement using beamforming and nonstationarity with applications to speech”, IEEE Trans. Signal Process., Vol. 49, No. 8, pp. 1614-1626.

[6] Jung, Y. et al. (2005). “Adaptive microphone array system with two-stage adaptation mode controller”, IEICE Trans. Fund., Vol. E88-A, No. 4, pp. 972-977.

[7] Hyvarinen, A. and Oja, E. (2000). Independent component analysis: Algorithms and applications, Neural Networks, vol. 13, no. 4, pp. 411-430.

[8] Hayes, M. (1996). Statistical Digital Signal Processing and Modeling, John Wiley & Sons.

[9] Rabiner, L. R. and Juang, B. H. (1993). Fundamentals of Speech Recognitions, Prentice Hall.

[10] *www.sitec.or.kr*

[11] http://htk.eng.cam.ac.uk

• **이영재 (Lee, Youngjae)**
경상대학교 전자공학과
경남 진주시 가좌동 900 번지

Tel: 055-751-5357
Email: clever1999@gnu.ac.kr
관심분야: 음성신호처리
현재 전자공학과 학부 재학중

• **김수환(Kim, Suhwan)**
경상대학교 전자공학과
경남 진주시 가좌동 900 번지
Tel: 055-751-5357
Email: edps2116@gnu.ac.kr
관심분야: 음성신호처리
현재 전자공학과 학부 재학중

• **한승호(Han, Seungho)**
한국과학기술원 정보통신공학과
대전광역시 유성구 문지로 119
Tel: 042-350-6206
Email: space0128@kaist.ac.kr
관심분야: 음성신호처리
현재 정보통신공학과 대학원 박사과정 재학중

• **한민수(Hahn, Minsoo)**
한국과학기술원 정보통신공학과
대전광역시 유성구 문지로 119
Tel: 042-350-6123
Email: mshahn@kaist.ac.kr
관심분야: 음성/음향신호처리, 디지털미디어
1998~현재 정보통신공학과 교수

• **김영일(Kim, Young-Il)**
경상대학교 전자공학과(공학연구원)
경남 진주시 가좌동 900 번지
Tel: 055-751-5352
Email: yi@gnu.ac.kr
관심분야: 음성/음향신호처리
1987~현재 전자공학과 교수

• **정상배(Jeong, Sangbae) 교신저자**
경상대학교 전자공학과(공학연구원)
경남 진주시 가좌동 900 번지
Tel: 055-751-5357
Email: jeongsb@gnu.ac.kr
관심분야: 음성/음향신호처리
2009~현재 전자공학과 조교수