칼만필터 기반의 다채널 입출력 음향학적 반향제거 방법에 관한 연구

박지환*, 장준혁**
*한양대학교 전자컴퓨터통신공학과
**한양대학교 융합전자공학부
e-mail:jchang@hanyang.ac.kr

A Study on MIMO Acoustic Echo Cancellation Based on Kalman filtering

Jihwan Park*, Joon-Hyuk Chang**
*Dept of Electronics and Computer Engineering, Hanyang University
**School of Electronic Engineering, Hanyang University

요 약

본 논문에서는 기존의 단일입출력 환경에서의 칼만필터 기반 반향제거방법을 다중입출력 구조로 확장하는 방법을 제안한다. 다중입출력 구조의 반향제거방법은 단일입출력방식보다 우수한 반향제거 성능을 보이면서도 더욱 낮은 음성왜곡도를 보였다.

1. 서론

스피커와 마이크가 한 공간에 존재할 때, 스피커에서 출력되는 신호가 마이크로 입력되는 현상이 발생하는데, 이때 마이크로 입력되는 신호를 음향학적 반향신호(acoustic echo)라 한다. 이러한 반향신호는 음성통신 감도를 떨어뜨리고, 음성인식율을 저하시키는 요인으로 작용하기 때문에, 음향학적 반향신호 제거 방법(acoustic echo cancellation)을 이용해 음성통신 및 인식 성능을 시킬 수 있다. 반향신호 제거는 반향신호가 입력되는 마이크 개수에 따라 그 방법에 차이가 있다[1]. 단일채널 마이크 기반의 반향신호 제거는 least mean square (LMS), recursive least squares (RLS) 등의 적응형 필터를 이용해 반향신호를 추정하고 이를 마이크 입력신호로부터 제거한다[2]. 다채널 마이크 기반의 반향신호 제거는 주로 빔포밍과 결합한 형태의 방법들이 많이 연구되었다[2]. LMS, RLS와 같은 적응형 필터를 이용해 각 마이크 채널별 반향신호 제거 후 빔포밍을 이용해 단일채널 출력을 하는 방법, 빔포밍을 거친 단일채널 출력에 대해 반향제거 하는 방법이 주를 이루고 있다. 해당 방법들은 공간적 필터링 능력이 우수한 빔포밍과의 결합을 통해 단일채널 반향제거 방법보다는 그 성능이 우수하나, 다음과 같이 2가지 단점이 존재한다. 첫째로는 출력신호가 단일채널이기 때문에 반향신호 제거 이후의 공간상의 정보를 얻을 수 없어, 반향신호를 제거한 이후의 유의미한 방향추정 및 다채널 마이크 기반의 상황인지를 수행할 수 없다. 둘째로는, 고정된 step size 값을 갖는 적응형 필터가 화자의 움직임에도 쉽게 변하는 반향환경에 빠르게 적응하지 못함에 따른 반향제거 성능 저하가 발생한다. 이에 대한 해결책으로, 반향환경에 대해 동적모델링을 함으로써 매 프레임 최적의 step size를 계산하는 칼만필터(Kalman filter) 기반의 반향제거가 연구되고 있다[3].

본 논문에서는 다채널 입출력 기반의 반향제거 방법을 제안한다. 다채널 입출력이 가능하도록 입력과 출력신호를 선형적 관계로 표현하고, 다채널 입출력 선형적 관계를 칼만필터를 이용해 동적모델링함으로써 반향환경에 빠르게 적응할 수 있도록 구성하였다. 시뮬레이션을 통해, 제안하는 방법이 기존의 단일채널 마이크 기반의 반향제거 방법보다 성능이 우수함을 확인하였다.

2. 다채널 입출력 환경에서의 반향제거 방법

다채널 마이크로 반향신호가 입력되는 환경에서 short-time Fourier transform (STFT) 영역에서의 다채널 마이크 입력신호를 아래와 같이 정의할 수 있다.

$$\mathbf{y}(k,n) = \sum_{l=0}^{L-1} \mathbf{g}_l(k,n) \bullet X(k,n) + \mathbf{s}(k,n) \qquad (1)$$

여기서 k는 STFT 주파수 인덱스, n은 프레임 인덱스, $\mathbf{y}(k,n) = [Y_1(k,n), ..., Y_M(k,n)]^T$는 다채널 마이크 입력신호 벡터로써, $Y_m(k,n)$은 m채널 마이크 입력의 STFT 계수를 의미하고, M은 전체 마이크 개수를 나타낸다. $\mathbf{s}(k,n)$는 다채널 음성신호 벡터이며, $\mathbf{y}(k,n)$처럼 정의된다. $X(k,n)$은 스피커 출력되기 전의 근단 신호의 STFT 계수이며, $\mathbf{g}_l(k,n)$는 필터차수 L의 다채널 음향학적 전달함수이다. 수식 (1)에서 $\mathbf{g}_l(k,n)$를 추정함으로써 반향신호 추정치를 계산할 수 있고, 반향제거 신호를 얻을 수 있다.

$$\hat{\mathbf{s}}(k,n) = \mathbf{y}(k,n) - \left[\sum_{l=0}^{L-1} \hat{\mathbf{g}_l}(k,n) \bullet X(k,n)\right] \quad (2)$$

3. 다채널 입출력 환경에서의 칼만필터를 이용한 동적 모델링 및 칼만필터 업데이트 방법

수식 (2)에서 설명한 바와 같이, 다채널 음향학적 전달함수를 추정함으로써 다채널 입출력 반향제거 방법이 동작가능하다. 동적모델링을 위해 다채널 음향학적 전달함수를 아래와 같이 1차 마코브 모델로 표현할 수 있다.

$$\mathbf{g}_l(k,n) = A \bullet \mathbf{g}_l(k,n-1) + \mathbf{w}(k,n) \quad (3)$$

여기서 A는 $\mathbf{g}_l(k,n)$의 프레임간 변이 상수이며, $\mathbf{w}(k,n)$는 프로세스 노이즈로써 평균이 0인 복소 가우시안으로 정의된다. 이러한 동적 모델링된 $\mathbf{g}_l(k,n)$를 효과적으로 추정하기 위해서 칼만필터를 이용한다[4].

$$\mathbf{g}^{+}(k,n) = A \bullet \mathbf{g}(k,n-1) \quad (4)$$

$$\boldsymbol{\Phi}_{\mathrm{w}}^{+}(k,n) = A^2 \bullet \hat{\boldsymbol{\Phi}}_{\mathrm{w}}(k,n-1) + \boldsymbol{\Phi}_{\Delta \mathrm{w}} \quad (5)$$

$$\mathbf{K}(k,n) = \boldsymbol{\Phi}_{\mathrm{w}}^{+}(k,n)\mathbf{D}^{H}(k,n) \bullet \left[\mathbf{D}(k,n)\boldsymbol{\Phi}_{\mathrm{w}}^{+}(k,n)\mathbf{D}^{H}(k,n) + \boldsymbol{\Phi}_{\mathrm{s}}(k,n)\right]^{-1} \quad (6)$$

$$\hat{\mathbf{g}}(k,n) = \mathbf{g}^{+}(k,n) + \mathbf{K}(k,n) \bullet \left[\mathbf{y}(k,n) - \mathbf{D}(k,n)\mathbf{g}^{+}(k,n)\right] \quad (7)$$

$$\hat{\boldsymbol{\Phi}}_{\mathrm{w}}(k,n) = \left[\boldsymbol{I} - \mathbf{K}(k,n)\mathbf{D}(k,n)\right] \bullet \boldsymbol{\Phi}_{\mathrm{w}}^{+}(\mathrm{k,n}) \quad (8)$$

여기서 $^{+}$는 칼만필터의 예측단계를 의미한다. $\boldsymbol{\Phi}_{\Delta \mathrm{w}}(\mathrm{k,n})$는 프로세스 노이즈의 공분산을 의미한다. 또한, $\mathbf{D}(k,n)$과 $\mathbf{g}(k,n)$는 아래와 같이 정의한다.

$$\mathbf{D}(k,n) = \boldsymbol{I} \otimes [X(k,n), ..., X(k,n-L+1)] \quad (9)$$

$$\mathbf{g}(k,n) = \left[\mathbf{g}_0^{H}(k,n), ..., \mathbf{g}_{L-1}^{H}(k,n)\right]^{H} \quad (10)$$

$\otimes$는 크로니컬 곱을 의미한다. 칼만필터를 이용해 추정된 다채널 음향학적 전달함수는 수식 (2)를 이용해 최종 다채널 출력신호를 계산할 수 있다.

4. 시뮬레이션

성능 측정을 위한 실험은 시뮬레이션 환경에서 수행되었다. $6 \times 6 \times 10\,m^3$의 공간에서 잔향시간 0.3초, 0.5초에 대한 임펄스 함수를 생성하여 시뮬레이션을 수행하였다[5]. 마이크는 2, 4개를 사용하였으며, 마이크 사이의 간격은 8 cm로 원형구조를 가정하였다. 화자는 마이크로부터 1m 떨어져있으며, 반향신호 재생을 위한 스피커는 마이크로부터 0.5m의 거리에 위치시켰다. 화자와 스피커는 90도의 각도차로 배치되었다. 화자신호는 무향실에서 녹음된 남성화자 음성을, 반향신호는 무향실에서 녹음된 여성 화자의 음성을 사용하였다. 기존의 방법[5]과 제안된 방법 모두 1024 크기의 STFT를 사용하였으며, 75% 오버랩이 이용되었다. 또한 필터차수 L은 1의 값을 사용하였다.

반향제거 성능검증을 위해 반향제거정도를 측정하는 ERLE[5]와 음성왜곡정도를 측정하는 PESQ[6]를 이용하였다. 2가지 잔향환경에서 마이크 개수에 증가에 따라 ERLE와 PESQ가 개선되는 양상을 보였다. 이를 통해 마이크 개수 증가에 따라 반향제거 방법의 성능이 개선되는 것을 확인할 수 있었다.

<표1 마이크개수에 따른 반향제거 성능 평가>

	$T_{60} = 0.3\,s$		$T_{60} = 0.5\,s$	
	ERLE (dB)	PESQ	ERLE (dB)	PESQ
$M=1$	13.88	1.79	8.77	1.46
$M=2$	17.05	2.09	10.04	1.57
$M=4$	24.27	2.77	14.84	1.96

5. 결론

본 논문에서는 칼만필터 기반의 다채널 입출력 반향제거 방법을 제안하였다. 단일채널 대비 다채널 마이크 환경에서의 반향신호가 더욱 정교하게 추정되었으며, 기존 반향제거 방법 대비 음성왜곡을 줄이면서도 반향제거 성능을 향상시킬 수 있었다.

감사의 글

이 논문은 2017년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No. 2016-0-00564, 사용자의 의도와 맥락을 이해하는 지능형 인터랙션 기술 연구개발)

참고문헌

[1] S. Haykin, Adaptive Filter Theory, Second Edition, EnglewoodCliffs, NJ, USA: Prentice-Hall, Sep. 1993.

[2] W. Herbordtt, S. Nakamura and W. Kellermann, "Joint optimization of LCMV beamforming and acoustic echo cancellation for automatic speech recognition," Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005, pp. iii/77-iii/80, 2005.

[3] G. Enzner and P. Vary, "Frequency-domain adaptive Kalman filter for acoustic echo control in hands-free telephones." *Signal Processing*, Vol. 86, No. 6, pp. 1140-1156, Jun., 2006.

[4] R. E. Kalman, "A new approach to linear filtering and prediction problems," *ASME Journal of Basic Engineeing*, Vol. 82, pp. 35-45, 1960.

[5] S. Malik and G. Enzner, "Recursive Bayesian control of multichannel acoustic echo cancellation." *IEEE Signal Processing Letters*, Vol. 18, No. 11, pp. 619-622, Nov. 2011.

[6] S. Y. Lee and N. S. Kim, "A statistical model based residual echo suppression," *IEEE Signal Process. Lett.*, Vol. 14, No. 10, pp. 758 - 761, Oct. 2007.

[7] "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs", ITU-T Rec., pp. 862, 2000.