시간-주파수 마스킹과 고차 신호 통계를 이용한 음향 반향신호 제거

論 文
56-3-29

Acoustic Echo Cancellation using Time-Frequency Masking and Higher-order Statistics

金 景 在* · 南 尙 沅†
(Kyoung-Jae Kim · Sang-Won Nam)

Abstract - In hands-free full-duplex communication systems, acoustic signals picked up by the microphones can be mixed with echo signals as well as noises, which may result in poor performance of the corresponding communication system. Also, the system performance may decrease further if the reverberation occurs since it is harder to estimate the impulse response of the demixing system. For blind source separation (BSS) in such cases, a time-frequency masking approach can be employed to separate undesired echo signals and noises, but, permutation ambiguities also should be solved for the echo cancellation. In this paper, we propose a new acoustic echo cancellation (AEC) approach utilizing the time-frequency masking and higher-order statistics, whereby a desired signal selection, based on coherence and third-order statistics (i.e., kurtosis), is introduced along with output signal normalization. Simulation results demonstrate that the proposed approach yields better echo and noise cancellation performances than the conventional AEC approaches.

Key Words : Acoustic Echo Cancellation, Blind Source Separation, Permutation Ambiguities, Time-frequency Masking

1. 서 론

Acoustic echo cancellation (AEC : 음향 반향신호 제거) 기법은 hands-free cellular telephony, internet telephony, audio나 video conferencing과 같은 시스템에서 적응 필터링에 기반하여 반향신호 제거에 효과적으로 사용되어왔다[1]. 특히 반향신호는 hands-free와 같은 양방향 통신 시스템에서 스피커와 마이크 사이의 open-air 경로에서 발생될 수 있다. 일반적으로 음향 반향신호 제거를 위해 기존의 적응 신호처리 기법을 적용하여 반향경로를 추적하는 방식을 취하고 있는데, 소음이 많은 상황에서는 성능이 낮아질 수 있는 단점을 갖고 있다[2]. 또한, 잔향(reverberation)이 있는 환경에서 기존의 음향 반향신호 제거 알고리즘에서는 긴 필터 계수 길이를 갖는 demixing 필터가 요구되어 필터 계수 추정에도 많은 어려움이 있다[2]. 따라서, 실제 상황에서 반향신호와 잔향과 더불어 소음까지 제거하기 위한 새로운 기법이 요구되어 왔다.

이러한 문제를 해결하기 위하여 본 논문에서는 시간-주파수 마스킹 (time-frequency masking) 방법에 기반한 암묵신호분리 (blind source separation: BSS) 방법과 고차통계를 이용하여 잔향 및 소음 환경에서 음향 반향신호를 효과적으로 분리하고 주요 원천 신호 성분을 추출하는 방법을 제안한다. 암묵신호분리는 원천 신호와 mixing 시스템에 대한 정보가 없더라도 섞여 들어온 신호만으로 원천 신호로 분리해 내는 방법으로, 과거 암묵신호분리 방법이 음향 반향신호 제거 방법에 사용된 적이 있으나 원천 신호의 수와 마이크로폰으로 측정된 신호의 수가 같아야만 하는 한계 상황에서 적용되었고, 잔향이 없고 소음이 섞이지 않은 이상적인 상황에서 고려되었으나, 잔향환경에서 소음까지 섞인 환경에서는 신호분리 성능이 현저히 떨어지기 때문에 암묵신호분리 기법에서의 순열(permutation) 등 실제적인 문제 해결이 어려웠다[3,4].

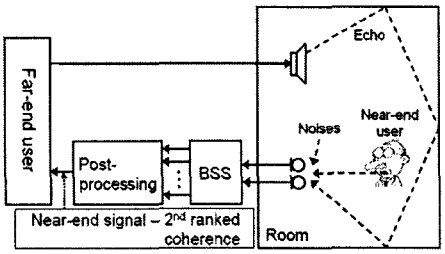

그림 1 음향 반향신호 제거방법 블록 다이어그램
Fig. 1 Block diagram for AEC

본 논문에서는 음성신호에 최적화된 암묵신호분리 기반의 시간-주파수 마스킹 (time-frequency masking) 방법[7]을 적용하고, 암묵신호분리 방법에서의 순열 문제를 해결하기 위해 고차통계인 kurtosis[8]와 coherence[9] 값들을 활용하여, 실제 상황(소음과 잔향환경)에서도 우수한 성능을 보이는 효율적 음향 반향신호 제거 알고리즘을 제안한다. 특히, 원천 신호들이 음성신호일 경우 시간-주파수 영역에서 관찰하면 거의 서로 중복되지 않는 W-disjoint orthogonality (W-DO) 특징[5]를 갖고 있고, 음성 신호의 W-DO 특징은 잔향환경에서도 그대로 유지되는 성질[6]을 이용하여 기존의 제한적인 방법을 개선하였다.(그림 2 및 그림 3 참조). 다음 2장에서는 새로운 음향 반향신호 제거 방법을 설명하고, 3장에서는 제안된 방법의 성능을 모의실험을 통해 확인하며, 4장에서 결론을 맺는다.

2. 새로운 음향 반향신호 제거 방법

2.1. 전체 암묵신호분리 구조

잔향환경에서 소음이 섞였을 때 효과적으로 음향 반향신호를 제거하기 위해 사용된 시간-주파수 마스킹 방법[5~7]은 실제 환경과 같은 underdetermined 상황(원천신호가 획득된 섞인 신호보다 많은 경우)에서도 우수한 성능을 보인다. 특히, 시간-주파수

† 교신저자, 正會員 : 漢陽大學校 電子通信컴퓨터學科 敎授 · 工博
E-mail : swnam@hanyang.ac.kr
* 學生會員 : 漢陽大學校 電子通信컴퓨터學科 碩士課程
接受日字 : 2007年 1月 5日
最終完了 : 2007年 2月 8日

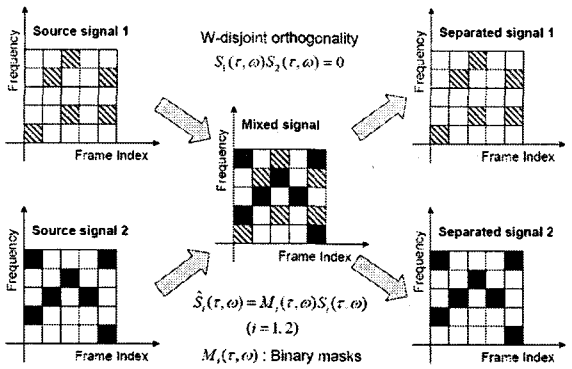

그림 2 시간-주파수 마스킹 방법을 적용한 음성신호분리
Fig. 2 Speech signal separation using the time-frequency masking[7]

마스킹을 사용한 알고리즘을 이용하여 섞인 신호를 분리하면 단지 두 개의 센서만을 이용하여도 기존의 음향 반향신호 제거 방법보다 개선된 성능을 얻을 수 있다. 즉, 잔향이 많아질수록 성능이 떨어지는 기존 방법과 달리 잔향 환경에서도 음성의 W-DO가 만족되므로 높은 신호분리 성능을 보인다. 본 논문에서 제안하는 새로운 음향 반향신호 제거 방법에서 가장 핵심적인 부분은 원하는 신호와 소음을 어떻게 정확히 구분해 내는 가이다. 즉, 암묵신호분리 알고리즘 자체가 가지는 순열(permutation) 문제는 음향 반향신호 제거 시 큰 걸림돌이 되어왔는데, 본 논문에서는 그 문제를 해결하기 위하여 스케일 추정과 더불어, 신호의 고차통계인 kurtosis와 coherence를 사용한다. 제안된 음향 반향신호 제거에 대한 블록 다이어그램인 그림 1의 구조를 보면 시간-주파수 마스킹이 사용된 암묵신호분리(BSS) 방법과 원하는 최종 음성 신호를 얻기 위한 후처리 과정(postprocessing)을 통해 잔향환경에서도 소음이 제거된 near-end 신호를 얻어낼 수 있도록 하였다.

2.2. 암묵신호분리(BSS)

암묵신호분리 알고리즘에서는 matrix inversion demixing 문제로 인하여 원천 신호의 수와 섞여 들어온 신호의 수 (또는 센서의 수)가 동일한 것이 일반적이다. 그러나, 이러한 가정은 실제 환경에서는 원천 신호의 수가 많을 때 그만큼 mixing 채널을 준비하는 것 자체가 어려운 상황이 발생할 수도 있다. 최근에 제안된 이진(binary) 시간-주파수 마스킹(time-freaquency masking)을 이용한 BSS 알고리즘[7]은 주어진 원천 신호가 다음의 W-DO를 만족할 때, 단지 두 개의 센서 신호만으로도 임의의 원천 신호들을 모두 분리해낼 수 있는 빠르고 효과적인 방법이다.

$$S_1(\tau,\omega)S_2(\tau,\omega)=0,\quad \forall\tau,\omega. \tag{1}$$

(1)에서 $S_1(\tau,\omega)$과 $S_2(\tau,\omega)$는 원천신호 $s_1(t)$과 $s_2(t)$을 각각 short-time Fourier transform(STFT)한 신호이다. 따라서, 이 특성을 적절히 이용하면 시간-주파수 영역에서 sparseness를 갖는 음성 신호의 경우 신호분리에 매우 효과적으로 적용 가능하다. 본 논문에서 사용한 시간-주파수 마스크는 [7]에서 사용한 방법을 따랐다(그림 2 참조). 또한, 잔향환경에서 녹음된 음성신호라도 W-DO를 만족하기 때문에 기존의 방법들이 제한적인 상황에서 적용되었던 것과 달리 실제 환경에서도 신호 분리가 가능하다[6]. 그림 3에 시간-주파수 영역에서 음성 신호의 W-DO 검증 예를 보였다. 즉, 실제 여성 및 남성 음성신호의 STFT를 그림 3(a) 및 그림 3(b)에 나타냈고, 시간-주파수 마스킹 기반 BSS를 이용하여 추정된 신호 (a) 및 신호 (b)의 추정 신호를 그림 3(c) 및 그림 3(d)에 나타내었다. 또한 그림 3(e)에 신호 (a)와 신호 (b)의 W-DO test 결과를 보였고, 그림 3(f)에 추정 신호 (c)와 신호 (d)의 W-DO test 결과를 보임으로써,

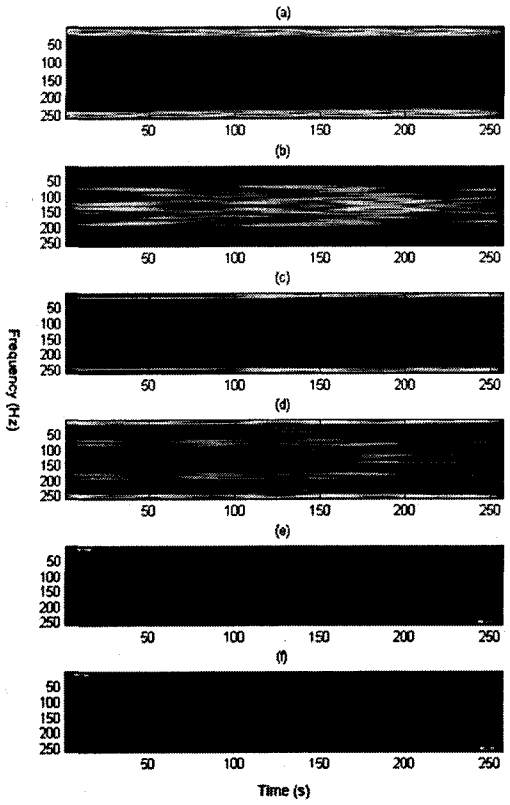

그림 3 시간-주파수 영역에서 음성 신호의 W-DO 검증: (a) 여성 음성신호, (b) 남성 음성신호, (c) 모의 잔향환경에서 얻어진 신호 (a)의 추정 신호, (d) 모의 잔향환경에서 얻어진 신호 (b)의 추정 신호, (e) 신호 (a)와 신호 (b)의 W-DO test, (f) 신호 (c)와 신호 (d)의 W-DO test (잔향환경에서도 W-DO가 만족함을 확인)
Fig. 3 Test for W-DO in the time-frequency domain.

잔향환경에서도 W-DO가 만족함을 확인하였다. 따라서 시간-주파수 마스킹을 사용한 알고리즘을 음향 반향신호 제거 기법에 적용하면 소음이 많은 잔향 환경에서도 원하는 음성 신호들을 효과적으로 분리해낼 수가 있다. 신호 분리 시 attenuation과 arrival delay의 두 가지 파라미터가 시간-주파수 영역에서 원천신호의 형태로 음성 신호 추정에 사용된다[7].

2.3. 원하는 음성 신호를 얻기 위한 후처리 과정

BSS 알고리즘은 원천 신호와 mixing 시스템에 대한 정보가 부족하므로 순열 문제가 항상 발생한다. 그러나 순열 문제는 주요 음성 신호를 추출하는 방법을 적용함으로써 해결될 수 있다. 이를 위해 본 논문에서는 nonGaussianity의 척도가 되는 kurtosis를 사용하였는데, Gaussian 신호의 경우 kurtosis 값은 0이고, super-gaussian 신호의 경우에는 양의 값, sub-gaussian 신호의 경우에는 음의 값을 보인다. 특히, 주요 음성 신호의 경우 super-gaussian 형태를 가지기 때문에 높은 kurtosis 값을 갖는 신호 순서대로 배열하여 쉽게 신호 구분을 할 수 있다. 환경 소음과 기기 잡음의 경우 주요 음성 신호보다 kurtosis 값이 작으므로 kurtosis의 크기대로 배열하였을 때 kurtosis가 가장 큰 두 개의 신호가 주요 음성 신호임을 쉽게 알 수 있다. 그러나, 그림 1에서와 같이 선택된 두 개의 음성 신호는 비슷한 kurtosis값을 가지므로, 어떤 신호가 near-end 신호이고 어떤 신호가 far-end 신호인지는 kurtosis의 값만으로는 구분해낼 수가 없다. 이를 위해 BSS를 통해 분리된 신호 중 첫째 및 둘째 크기의 kurtosis 값을 보이는 두 음성신호들과 far-end line 신호간의 coherence를 구하여, 작은 coherence 값을 보이는 신호가 원하는 near-end 신호임을 알 수 있다. 이는 분리된 신호 중

반향신호는 far-end line 반향신호와 더 큰 coherence 값을 보이기 때문이다.

3. 모의실험

모의실험을 위해 필요한 환경 소음과 모바일 기기에서 생기는 잡음은 http://freesound.iua.upf.edu/index.php에서 얻었고, 남성음성(near-end 신호)과 여성음성(far-end 신호)은 [7]에서 사용된 표본 음성 신호를 사용하였다. 이 신호들은 모두 16KHz로 표본화되어 있으며, 여성음성에는 일반적인 반향 경로를 적용하였다. 모의 잔향환경을 만들기 위해서 Cool Edit Pro 2.1의 delay effect중 shower room reverb를 사용하였다. 그림 4는 4개의 원천 신호로 이루어져 있으며, 왼쪽((a)~(d))은 원천신호들이고 오른쪽((e)~(h))은 분리된 신호들이다. Kurtosis와 coherence를 통하여 near-end 신호가 가장 위에 위치해 있음을 알 수 있다. 따라서 본 논문에서 제안된 알고리즘을 통해 소음이 제거된 near-end 신호(Fig. 4(e))와 far-end 신호(Fig. 4(f))를 확인할 수 있다. 또한, 분리된 신호들의 kurtosis와 coherence의 값들은 표 I과 같다. 소음이 많은 상황에서는 성능이 현저히 낮아지고, 잔향(reverberation)이 있는 환경에서 적응 필터 계수 추정에 많은 어려움을 겪는 기존 방법[2]에 비해 본 논문에서 제안한 방법은 효과적으로 near-end 신호를 분리해냄을 확인할 수 있다(Fig. 4(e)).

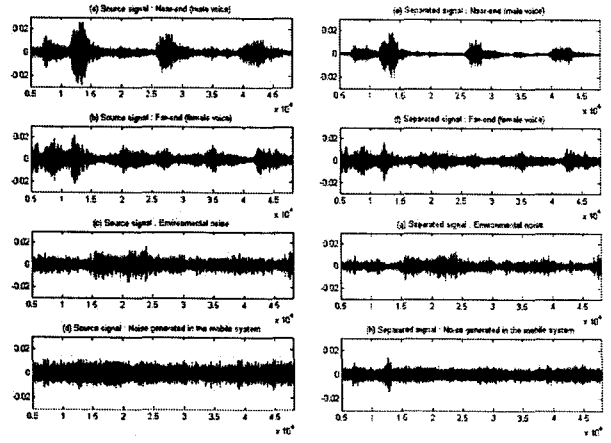

그림 4 원천신호 (a)~(d)와 분리된 신호 (e)~(h)

Fig. 4 Source signals (a)~(d) and separated signals (e)~(h)

표 1 분리된 신호의 kurtosis와 coherence 값

Table 1 Kurtosis and coherence values of separated signals

	Fig. 4(e)	Fig. 4(f)	Fig. 4(g)	Fig. 4(h)
Kurtosis	15.46	16.02	9.95	8.72
Coherence	0.11	0.45	·	·

4. 결 론

본 논문에서는 소음 및 잔향 환경에서 개선된 시간-주파수 마스킹에 기반을 둔 암묵신호분리 알고리즘과 고차통계를 이용하여 음향 반향신호를 제거하는 새로운 방법을 제안하였다. 음성 신호가 잔향환경에서도 W-DO를 만족한다는 조건[7]을 이용하여 음향 반향신호 제거 시 신호분리 성능을 높였고, 암묵신호분리 방법 자체가 갖는 순열 문제 해결을 위해 kurtosis와 coherence를 적용한 효율적인 음향 반향신호 알고리즘을 제안하였고, 모의실험을 통해 그 성능을 검증하였다.

Acknowledgement: This study was supported by a grant of the Korea Health 21 R & D Project, Ministry of Health & Welfare, Republic of Korea (02-PJ3-PG6-EV08-0001).

참 고 문 헌

[1] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*, New Jersey, Wiley, 2004.

[2] S. Haykin, *Adaptive Filter Theory*, 4th Ed., New Jersey, Prentice-Hall, 2002.

[3] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.

[4] D. Kim, H. Choi, and H. Bae, "Acoustic echo cancellation using blind source separation," *Proc. IEEE Workshop on SIPS2003*, pp. 241-244, Aug. 2003.

[5] A. Jourjine, S. Rickard, and Ö. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures," *Proc. ICASSP2000*, vol. 5, pp. 2985-2988, Jun. 2000.

[6] S. Araki, S. Makino, A. Blin, R. Mukai, and H. Sawada, "Underdetermined blind separation for speech in real environments with sparseness and ICA," *Proc. ICASSP2004*, vol. 3, pp. 881-884, May 2004.

[7] Ö. Yılmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 52, no. 7, pp. 1830-1847, Jul. 2005.

[8] S.Y. Low, S. Nordholm, and R. Togneri, "Convolutive blind signal separation with post-processing," *IEEE Trans. on Speech and Audio Process.*, vol. 12, no. 5, pp. 539-548, Sept. 2004.

[9] T. Gänsler, M. Hansson, C.J. Ivarsson, and G. Salomonsson, "A double-talk detector based on coherence," *IEEE Trans. on Communications*, vol. 44, no. 11, pp. 1421-1427, Nov. 1996.