음성존재확률을 이용한 최적 변형 다채널 위너 필터

(An Optimally-Modified Multichannel Wiener Filter Using Speech Presence Probability)

정상배*, 김영일**

(Sangbae Jeong, Youngil Kim)

요 약

본 논문에서는 음성존재확률을 이용하여 다채널 위너필터의 이득을 최적으로 변형하는 방법을 제안한다. 기존의 음성존재확률을 이용한 다채널 위너필터의 변형은 다소 경험적인 방법을 사용하기 때문에 잔여잡음의 양을 줄이면 음성왜곡이 증가하는 문제가 있다. 하지만, 제안된 최적 변형 다채널 위너필터는 음성존재확률을 최적 필터를 도출하기 위한 비용함수에 적용하여 비제한적 최소화 문제의 해를 이용하여 잔여잡음의 양과 음성왜곡을 동시에 줄일 수 있는 결과를 보였다. 잡음제거된 파형과 스펙트로그램의 평가를 통해서 제안된 최적 변형 다채널 위너필터가 종래의 다채널 위너필터와 비교하여 향상된 SNR과 음성왜곡을 나타냄을 확인할 수 있었다.

■ 중심어 : 음성존재확률 ; 다채널 위너필터 ; 이득 변형

Abstract

This paper proposes an optimal gain modification method of the Multichannel Wiener filter (MWF) using speech presence probabilities. Conventional gain modification methods of MWFs have the problem of the increase of speech distortions while reducing residual noises with its relative heuristic approach. However, the proposed optimal gain modification method, derived by solving the unconstrained minimization problem of the probability-involved cost function, reduces amounts of residual noises and signal distortions simultaneously. Through an evaluation of the filtered waveforms and spectrograms, it is verified that the proposed method results in an improved SNR with less signal distortions compared to the conventional MWF.

■ keywords : speech presence probability ; multichannel Wiener filter ; gain modification

Ⅰ. 서 론

최근, 깊은 신경망 기술(deep nural network)의 학습방법 고도화에 따라서 음성기반의 서비스가 크게 발전하고 있다[1,2]. 이에, 음성을 열화 시키는 잡음의 제거 기술에도 관심이 높아지고 있다. 그 중, 위너필터 (Wiener filter)는 추정된 신호와 원 신호간의 차이, 즉 추정 오차를 최소화하는 최적 필터로서 정상성(staitonary) 잡음을 제거하는데 뛰어난 성능을 보인다[3]. 그러나 비정상성(nonstationary) 잡음을 제거하는데 취약하기 때문에 다채널 마이크로폰 배열 신호처리를 위해 다채널로 확장된 위너필터가 제안되었고, 다채널 위너필터(multichannel Wiener filter)는 방향성 잡음과 주변 잡음이 함께 존재하는 열악한 환경에서 기존 단채널 위너필터 보다 향상된 성능을 보인다[4]. 최근에, 다채널 위너필터를 일반화 시킨 매개변수 내장형 다채널 위너필터(parameterized multichannel Wiener filter)가 음성강화 를 위해 제안되었고, 이 방법은 주파수 영역 음성강화 기법으로 잡음제거 성능을 측정하는 중요한 두 가지 기준인 음성 왜곡과 잔여 잡음 간의 트레이드오프 관계를 조절할 수 있는 장점이 있다[4,5]. 매개변수 내장형 위너필터는 음성신호와 잡음의 전력스펙트럼 밀도(power spectral density)와 같은 2차 통계치를 추정하여 구할 수 있으며, 잘 알려진 minimum variance distortionless response (MVDR)도 매개변수 내장형 위너필터의 한 가지 특별한 경우라는 사실이 증명되었다[5].

음성존재확률(speech presence probability))을 활용하여 기존 필터들을 변형하여 성능을 향상시키려는 시도가 있어왔는데 minimum mean square error-short time spectral amplitude (MMSE-STSA) 필터에 확률을 곱하여 필터를 변

* 정회원, 경상대학교 전자공학과/공학연구원
** 정회원, 경상대학교 전자공학과/공학연구원
이 연구는 2017년도 경상대학교 연구년제연구교수 연구지원비에 의하여 수행되었음.

접수일자 : 2017년 11월 20일 게재확정일 : 2018년 09월 21일
수정일자 : 1차 2018년 09월 11일, 2차 2018년 09월 19일 교신저자 : 김영일, e-mail : yi@gnu.ac.kr

형시키는 방법과 수식유도를 통해 최적으로 변형시키는 방법이 대표적인 연구이다[6,7]. 또한, 다채널 위너필터에도 성능향상을 위하여 다채널기반 음성존재확률(multichannel speech presence probability)을 다채널 위너필터에 도입하였지만[8], 경험적 이득 변형 (heuristic gain modification)은 잔여 잡음을 줄일 수 있으나 음성왜곡도가 증가하는 결과를 보였다.

따라서, 본 논문에서는 음성존재확률을 이용하여 매개변수 내장형 위너필터의 이득을 최적으로 변형하는 방법을 제안하였다. 제안된 최적변형 다채널 위너필터는 확률 기반의 비용 함수를 비제한적 최적화 문제(unconstrained optimization problem)를 풀어서 유도할 수 있다. 실험 결과를 분석한 결과, 제안한 최적변형 다채널 위너필터는 확률에 따라 자동으로 이득이 조절되는 매개변수 내장형 위너필터이며 잔여 잡음과 음성왜곡을 동시에 줄이는 결과를 보였다.

Ⅱ. 문제 수식화

m번째 잡음 섞인 입력신호의 단구간 푸리에 변환 (short time Fourier transform)을 $Y_m(t,k)$ 이라고 하면, 수식 (1)과 같이 표현할 수 있다.

$$Y_m(t,k) = X_m(t,k) + N_m(t,k),\ m = 1,2,...,M \quad (1)$$

$X_m(t,k)$ 과 $N_m(t,k)$ 은 각각 m번째 마이크로폰의 음성과 잡음의 단구간 푸리에 변환이고, t, k, M은 각각 시간 인덱스, 주파수 인덱스, 마이크로폰 개수를 의미한다. 다채널 잡음 섞인 입력신호를 벡터-행렬 형태로 표기하면 수식 (2)와 같다.

$$\begin{aligned} \boldsymbol{y}(t,k) &= \boldsymbol{x}(t,k) + \boldsymbol{n}(t,k) \\ &= \boldsymbol{a}(k) \cdot S(t,k) + \boldsymbol{n}(t,k) \end{aligned} \quad (2)$$

$\boldsymbol{y}(t,k)$ 는 $[Y_1(t,k),\ Y_2(t,k),\cdots,\ Y_M(t,k)]^T$ 로 구성된 입력신호 스펙트럼 벡터이고, $\boldsymbol{x}(t,k)$ 는 $[X_1(t,k),\ X_2(t,k),\cdots,\ X_M(t,k)]^T$ 형태의 마이크로폰에 입력된 음성신호 스펙트럼 벡터이며, $\boldsymbol{n}(t,k)$는 $[N_1(t,k),\ N_2(t,k),\cdots,\ N_M(t,k)]^T$ 형태의 잡음 스펙트럼 벡터이다. 위첨자 T는 전치(transpose)를 의미하고, $S(t,k)$ 는 음원으로서의 음성 스펙트럼을 나타낸다. 또한, $\boldsymbol{a}$(k) = $[A_1(k),\ A_2(k),\cdots,\ A_M(k)]^T$ 는 음원으로부터 M개의 마이크로폰까지의 전달함수로 구성된 벡터이다.

수식 (2)로부터 각 주파수의 전체 추정 오차를 구하면 다음과 같다.

$$\begin{aligned} \boldsymbol{e}(t,k) &= \boldsymbol{x}(t,k) - \hat{\boldsymbol{x}}(t,k) \\ &= (\boldsymbol{I} - \boldsymbol{G}(t,k))^H \boldsymbol{x}(t,k) + \boldsymbol{G}^H(t,k)\boldsymbol{n}(t,k) \\ &= \boldsymbol{e_X}(t,k) + \boldsymbol{e_N}(t,k) \end{aligned} \quad (3)$$

$\hat{\boldsymbol{x}}(t,k) = \boldsymbol{G}(t,k)\boldsymbol{y}(t,k)$ 은 추정된 신호이다. $\boldsymbol{G}(t,k)$= $[\boldsymbol{g}_1(t,k),...,\boldsymbol{g}_M(t,k)]$ 는 $M \times M$ 행렬로 주어지는 음성강화 필터 이며 $\boldsymbol{g}_m(t,k)$는 m 번째 출력 채널을 얻기 위한 필터 벡터이다. $\boldsymbol{e_X}(t,k)$, $\boldsymbol{e_N}(t,k)$, 위첨자 H, ^ 는 각각 음성신호 왜곡, 잔여잡음, 에르미트(Hermitian) 연산자, 변수의 추정치를 의미한다. 수식 (3)을 이용하여, 전체 추정오차의 에너지를 구하면 수식 (4)와 같다.

$$\overline{\epsilon^2}(t,k) = \overline{\epsilon_X^2}(t,k) + \overline{\epsilon_N^2}(t,k) \quad (4)$$

$\overline{\epsilon_X^2}$, $\overline{\epsilon_N^2}$는 각각 음성신호 왜곡, 잔여잡음 에너지를 의미하며 수식 (5)와 수식 (6)으로 표현된다.

$$\overline{\epsilon_X^2} = tr(E[\boldsymbol{e_X}(t,k)\boldsymbol{e_X^H}(t,k)]) \quad (5)$$

$$\overline{\epsilon_N^2} = tr(E[\boldsymbol{e_N}(t,k)\boldsymbol{e_N^H}(t,k)]) \quad (6)$$

다채널 위너필터는 수식 (4)의 전체 오차를 비제한적 최소화 문제로 풀면 얻을 수 있고, 매개변수 내장형 위너필터는 수식 (5)의 신호왜곡 에너지를 임의의 문턱치 이하로 유지하면서 수식 (6)의 잔여 잡음 에너지를 최소화하는 제한적 최적화 문제 (constrained optimization problem)를 풀면 얻을 수 있다. 그러나 일반적으로 음성왜곡과 잔여 잡음 간에는 트레이드오프 관계가 있으므로 동시에 향상시킬 수 없다[4-8]. 즉, 잔여잡음을 줄이면 음성왜곡도가 증가하게 되는 문제가 있다.

Ⅲ. 제안한 최적 변형 다채널 위너필터

앞서 언급한 음성 존재 확률을 이용한 매개변수 내장형 위너필터의 경험적 이득 조정은 최적이 아니기 때문에 트레이드오프 관계에 있는 음성왜곡과 잔여잡음을 동시에 향상시킬 수가 없다[8]. 위 문제를 해결하기 위해서 본 논문에서는 확률 기반의 비용함수를 제안하고 이 비용함수를 비제한적 최소화 문제로 풀어서 매개변수 내장형 위너필터의 이득을 최적으로 조정하는 방법을 제안한다. 위에서 언급한 비제한적 최적화 문제는 수식 (7)과 같이 표현될 수 있다.

$$\min_{\boldsymbol{G}(t,k)} \overline{\epsilon_p^2}(t,k) \quad (7)$$

$$\overline{\epsilon_p^2} = p(t,k)\overline{\epsilon_X^2}(t,k) + (1 - p(t,k))\overline{\epsilon_N^2}(t,k) \quad (8)$$

수식 (8)은 확률기반의 비용함수이고 $p(t,k)$ 는 t번째 프레임의 k번째 주파수 성분의 다채널기반 음성존재확률로서 다채널 온라인 잡음 추정 및 추적 방법을 통해 구할 수 있다[8]. 수식 (8)의 확률기반의 비용함수는 다채널 기반 음성존재확률에 의해 자동적으로 조절될 수 있다. 즉, $p(t,k)$가 1에 가까울수록 음성이 존재할 확률이 높기 때문에 음성왜곡 에너지를 최소화 시키며, $p(t,k)$가 0에 가까울수록 음성이 존재하지 않을 확률이 높기 때문에 잔여잡음 에너지를 최소화 시킨다.

수식 (8)을 $\boldsymbol{G}^H(t,k)$에 대하여 미분하면, 제안한 최적 변형 다채널 위너필터의 n_0 번째 열벡터가 수식 (9)와 같이 구해진다.

$$\begin{aligned}\boldsymbol{g}_{n_0}^{(OM)}(t,k) &= [p(t,k)\boldsymbol{\Phi}_{\boldsymbol{XX}}(t,k)+(1-p(t,k))\boldsymbol{\Phi}_{\boldsymbol{NN}}(t,k)]^{-1} \\ &\quad \cdot p(t,k)\boldsymbol{\Phi}_{\boldsymbol{XX}}(t,k)\boldsymbol{u}_{n_0} \\ &= [\boldsymbol{\Phi}_{\boldsymbol{XX}}^{(SPP)}(t,k)+\boldsymbol{\Phi}_{\boldsymbol{NN}}^{(SAP)}(t,k)]^{-1}\boldsymbol{\Phi}_{\boldsymbol{XX}}^{(SPP)}(t,k)\boldsymbol{u}_{n_0}\end{aligned} \tag{9}$$

여기서, $\boldsymbol{u}_{n_0}=[0,,,,0,1^{n_0^{th}},0,...,0]^T$ 이며, n_0는 일반적으로 1로 설정한다. $\boldsymbol{\Phi}_{NN}(t,k) = E[\boldsymbol{n}(t,k)\boldsymbol{n}^H(t,k)]$, $\boldsymbol{\Phi}_{XX}(t,k) = E[\boldsymbol{x}(t,k)\boldsymbol{x}^H(t,k)]$, $\boldsymbol{\Phi}_{XX}^{(SPP)}(t,k)=p(t,k)\boldsymbol{\Phi}_{XX}(t,k)$, $\boldsymbol{\Phi}_{XX}^{(SAP)}(t,k)$ $=(1-p(t,k))\cdot\boldsymbol{\Phi}_{NN}(t,k)$ 는 각각 음성, 잡음, 음성 존재확률이 곱해진 음성, 음성부재확률이 곱해진 전력스펙트럼 밀도를 의미한다.

추가적으로, 제안한 최적 변형 다채널 위너필터는 내장형 위너필터와 같이 랭크(rank)가 1인 특징 ($\lambda(t,k)=\phi_{ss}(t,k)\boldsymbol{a}^H(k)\boldsymbol{a}(k)$, 이 때 $\phi_{ss}(t,k)$ 는 음성의 전력을 나타냄)과 woodbury 정리를 이용하여 보다 간단히 정리 될 수 있다[4,5]. 음성 존재확률 $p(t,k)$ 와 부재확률 $(1-p(t,k))$의 곱이 0이 아니라면, 수식 (10)이 성립한다.

$$\begin{aligned}&[p(t,k)\boldsymbol{\Phi}_{\boldsymbol{XX}}(t,k)+(1-p(t,k))\boldsymbol{\Phi}_{\boldsymbol{NN}}(t,k)]^{-1} \\ &= \frac{1}{(1-p(t,k))}\left[\boldsymbol{\Phi}_{\boldsymbol{NN}}^{-1}(t,k)-\frac{p(t,k)\boldsymbol{\Phi}_{NN}^{-1}(t,k)\boldsymbol{\Phi}_{XX}(t,k)\boldsymbol{\Phi}_{NN}^{-1}(t,k)}{1+(\lambda(t,k)-1)p(t,k)}\right]\end{aligned} \tag{10}$$

수식 (10)을 이용하여, 수식 (9)를 다시 표현하면 수식 (11)과 같다.

$$\begin{aligned}\boldsymbol{g}_{n_0}^{(OM)}(t,k) = \boldsymbol{I}_{\boldsymbol{M}} &- (1-p(t,k))[p(t,k)\boldsymbol{\Phi}_{\boldsymbol{XX}}(t,k) \\ &+(1-p(t,k))\boldsymbol{\Phi}_{\boldsymbol{NN}}(t,k)]^{-1}\end{aligned} \tag{11}$$

$\boldsymbol{I}_M$ 은 M 차원 항등 행렬이다. 수식 (11)을 정리하면, 수식 (12)와 같이 간소화된 최적변형 다채널 위너필터를 얻을 수 있다[5].

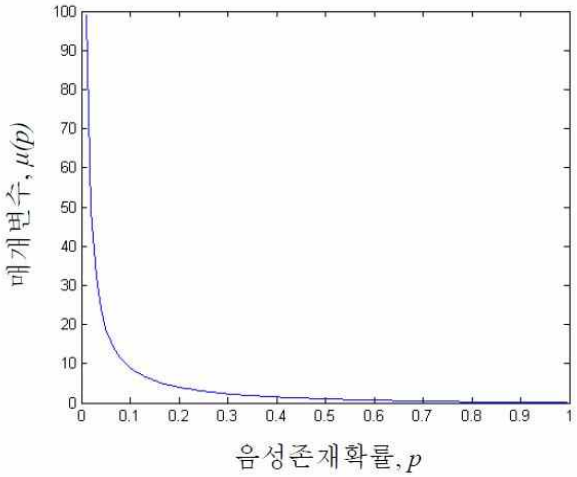

그림 1. 음성존재확률 값에 따른 제어함수

$$\begin{aligned}\boldsymbol{g}_{n_0}^{(OM)}(t,k) &= \frac{p(t,k)\boldsymbol{\Phi}_{\boldsymbol{NN}}^{-1}(t,k)\boldsymbol{\Phi}_{\boldsymbol{XX}}(t,k)}{1+(\lambda(t,k)-1)p(t,k)}\boldsymbol{u}_{n_0} \\ &= \frac{p(t,k)\boldsymbol{\Phi}_{NN}^{-1}(t,k)\boldsymbol{\Phi}_{YY}(t,k)-p(t,k)\boldsymbol{I}_M}{1+(\lambda(t,k)-1)p(t,k)}\boldsymbol{u}_{n_0}\end{aligned} \tag{12}$$

$\boldsymbol{\Phi}_{YY}(t,k)=E[\boldsymbol{y}(t,k)\boldsymbol{y}^{H(t,k)}]$ 은 잡음 섞인 입력의 전력스펙트럼 밀도 행렬이고, 음성과 잡음이 무상관(uncorrelated) 관계라고 가정하면, $\boldsymbol{\Phi}_{YY}(t,k)=\boldsymbol{\Phi}_{XX}(t,k)+\boldsymbol{\Phi}_{NN}(t,k)$ 인 관계가 성립한다. 또한, $p(t,k)\neq 0$ 을 가정하고 수식 (9)와 (12)를 $p(t,k)$ 로 각각 나누면, 수식 (13)과 (14)를 얻을 수 있다.

$$\begin{aligned}\boldsymbol{g}_{n_0}^{(OM)}(t,k) = &[\boldsymbol{\Phi}_{\boldsymbol{XX}}(t,k)+\mu(p(t,k))\boldsymbol{\Phi}_{\boldsymbol{NN}}(t,k)]^{-1}\cdot \\ &\boldsymbol{\Phi}_{\boldsymbol{XX}}(t,k)\boldsymbol{u}_{n_0}\end{aligned} \tag{13}$$

$$\boldsymbol{g}_{n_0}^{(OM)}(t,k) = \frac{\boldsymbol{\Phi}_{NN}^{-1}(t,k)\boldsymbol{\Phi}_{XX}(t,k)}{\mu(p(t,k))+\lambda(t,k)}\boldsymbol{u}_{n_0} \tag{14}$$

여기서, $\mu(p(t,k))=(1-p(t,k))/p(t,k)$ 는 다채널 기반 음성존재확률 값에 따라 변하는 제어 함수로서 그림 1과 같은 특성을 보인다. 수식 (13)과 (14)는 매개변수 내장형 위너필터[4,5]와 같은 형태이지만 음성왜곡도와 잔여잡음 간의 트레이드오프 관계를 음성존재확률에 따라 자동으로 조절할 수 있는 장점이 있다. 위 제어함수를 정리하면 수식 (15)를 얻을 수 있다.

$$\mu(p(t,k))=\frac{1-p(t,k)}{p(t,k)}=\frac{1-\frac{\Lambda(t,k)}{1+\Lambda(t,k)}}{\frac{\Lambda(t,k)}{1+\Lambda(t,k)}}=\Lambda^{-1}(t,k) \tag{15}$$

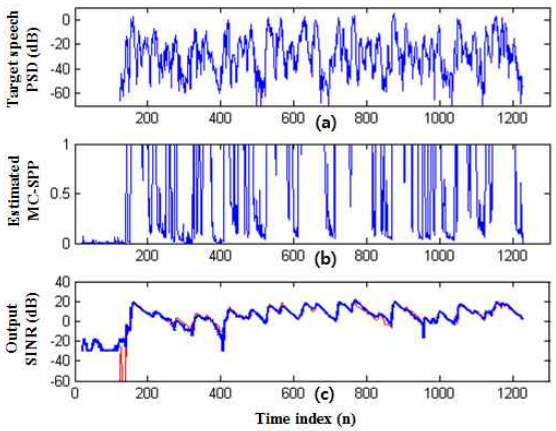

그림 2. 다채널 MCRA 잡음추정 및 추적 성능 검증: 배블잡음 (a) 음성 전력스펙트럼 밀도, (b) 추정된 다채널 기반 음성존재확률, (c) 추정된 출력 SINR (파란색) 과 이론적 출력 SINR (빨간색)

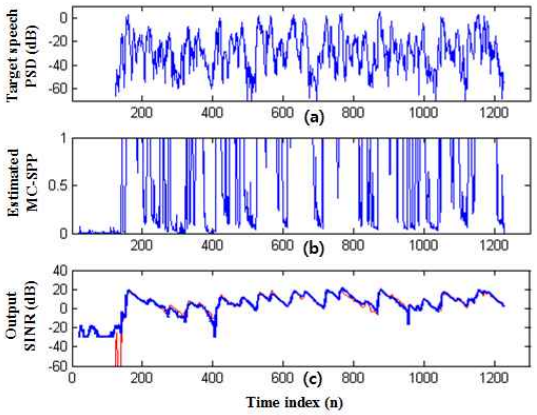

그림 3. 다채널 MCRA 잡음추정 및 추적 성능 검증: F-16 잡음 (a) 음성 전력스펙트럼 밀도, (b) 추정된 다채널 기반 음성존재확률, (c) 추정된 출력 SINR (파란색) 과 이론적 출력 SINR (빨간색)

여기서, $\Lambda(t,k)$는 t번째 프레임, k번째 주파수의 일반화된 우도비(generalized likelihood ratio)를 나타낸다[9].

IV. 실험 및 결과

실험의 전반적인 구성은 [8]의 조건과 같다. 모의된 공간(simulated chamber)의 크기는 3.048 m × 4.572 m × 3.810 m 였고, 음성과 방향성 간섭신호의 위치는 각각 (0.274 m, 3.181 m, 1.016 m) 와 (2.774 m, 3.181 m, 1.016 m) 였다. 마이크로폰 배열은 마이크 간 간격(d)가 0.069 m 인 균일 선형(uniform linear) 형태로 배치되었다. 첫 번째 마이크로폰 (기준 채널) 은 (1.282 m, 1.016 m, 1.016 m)에 배치되었고 j번째 마이크로폰은 (1.282 m + (j-1)d, 1.016 m, 1.016 m) 의 위치에 총 4개가 배치되었다. 반향시간 (T_{60})이 0 ms 인 무반향(anechoic) 환경과 T_{60} 이 210 ms 인 반향 환경을 고려하였다[8]. 하나의 깨끗한 음성 세트는 6개의 IEEE 문장들 [10]을 이어 붙여서 구성하였고, 총 10 세트의 깨끗한 음성 데이터베이스를 구축하였다. 총 60개의 IEEE 문장이 깨끗한 음성 데이터베이스로 사용되었다. 잡음 데이터베이스로는 NOISEX[11]의 "babble"과 "F-16"을 방향성 간섭 신호로 컴퓨터에서 생성된 가우시안 잡음을 주변잡음으로 추가하였다. 즉, 수식 (1)과 (2)의 m 번째 마이크로폰의 잡음 ($n_m(t)$) 은 m번째 마이크로폰의 방향성 잡음($i_m(t)$) 과 가산성 가우시안 잡음 ($w_m(t)$)의 합으로 구성되어 있다. 따라서, 신호 대 총 잡음 비(signal-to-interference-and-noise-ratio)는 SINR = $E[x_m^2(t)]/E[n_m^2(t)]$로 정의할 수 있으며, 비슷한 형태로, 신호 대 방향성 잡음비(signal-to-interference ratio)는 SIR = $E[x_m^2(t)]/E[i_m^2(t)]$로 신호대 주변 잡음비 (signal-to-ambient noise ratio)는 SNR = $E[x_m^2(t)]/E[w_m^2(t)]$로 각각 정의할 수 있다. 원거리 음성과 방향성 잡음을 생성하기 위해서 각각의 음원과 마이크로폰 간의 충격 응답(impulse response) 을 모의하기 위해서 이미지 기법(image method)을 사용하였고 [12,13], 각각의 음원은 해당되는 임펄스 응답과 합성곱(convolution)하여 잡음 섞인 입력신호를 생성하였다. 잡음섞인 다채널 입력신호 데이터베이스를 구축하기 위해서 깨끗한 음성, 방향성 잡음, 가산성 가우시안 잡음은 약 3.80과 8.81 dB의 SINR(SIR: 5, 15 dB, SNR: 10 dB)을 갖도록 인위적으로 합성하였다. 기존 방법 및 제안한 최적 변형 다채널 위너필터를 구현하기 위해, 프래임 길이는 무반향환경에서 32 ms를, 반향환경에서 64 ms로 각각 설정하였고, 50 % 의 분석 프래임 중첩(overlap)과 해밍 (Hamming) 윈도우를 사용하였다. 선형 필터들에서 발생할 수 있는 원형 합성곱 효과 (circular convolution effect)[14]을 방지하기 위해서 추정된 음성신호의 복원을 위해서 뒤쪽 절반의 신호를 이용하였다. 선형필터의 비인과성 (non-causality)을 보장하기 위해서, 세 가지 절차를 수행하였다: 필터를 시간영역으로 변환하고, 중간의 절반길이의 샘플 값을 취하여 다시 필터를 주파수 영역으로 변환하였다 [5,15,16]. 선형 필터링 후, 단구간 프래임은 오버랩-애드(overlap-add) 방식을 이용하여 복원하였다. 잡음의 추정 및 추적을 위해서 다채널 기반 minima controlled recursive averaging (MCRA) [8] 기법을 사용 하였고, 잡음 추정에 사용된 변수는 [8]의 값과 같게 설정하였다.

먼저, 구현된 다채널 기반 MCRA의 잡음 추정과 추적 성능을 검증하기 위해 그림 2와 그림 3과 같이 무반향환경 데이터베이스 중 1 kHz에서의 음성의 전력스펙트럼 밀도, 추정된 다채널 기반 음성존재확률, 출력 SINR을 각각 도시하였다. 그림 2는 방향성 잡음이 "babble"인 경우, 그림 3은 "F-16" 인 경우

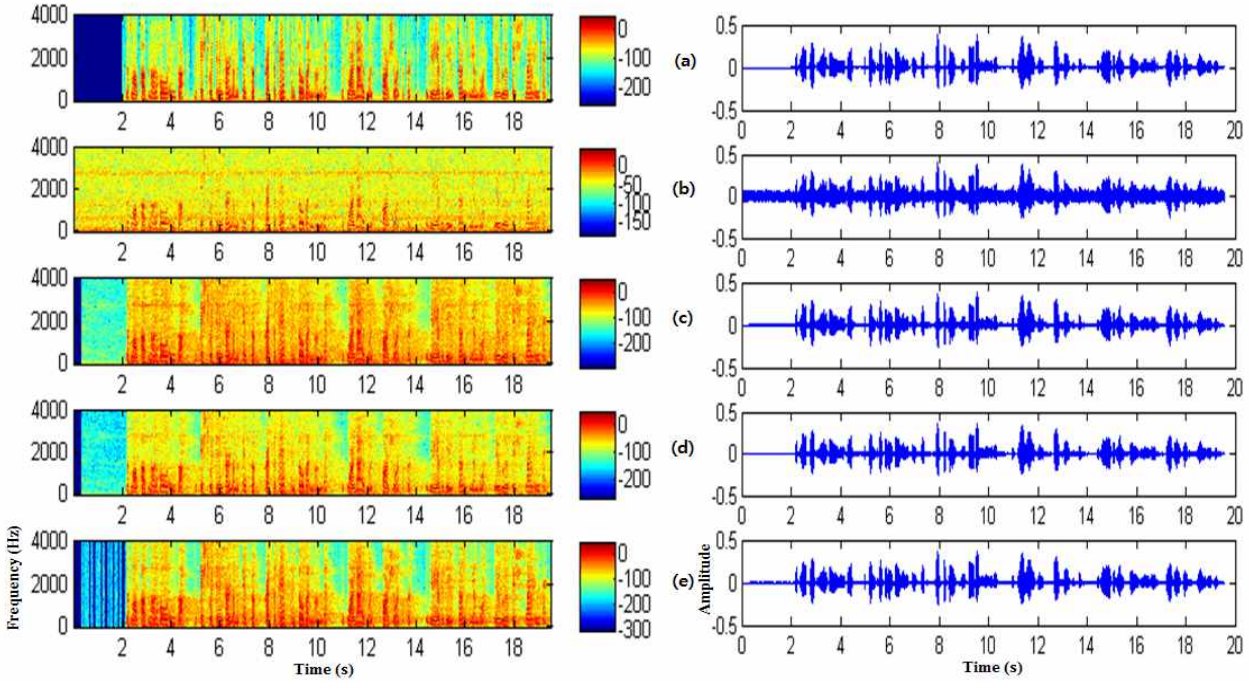

그림 4. 스펙트로그램 및 파형 결과 : F-16 잡음 (a) 깨끗한 음성, (b) 잡음 섞인 입력신호 (c) MVDR결과(β = 0) (d) 다채널 위너필터 결과(β = 1) (e) 제안한 최적 변형 다채널 위너필터 결과

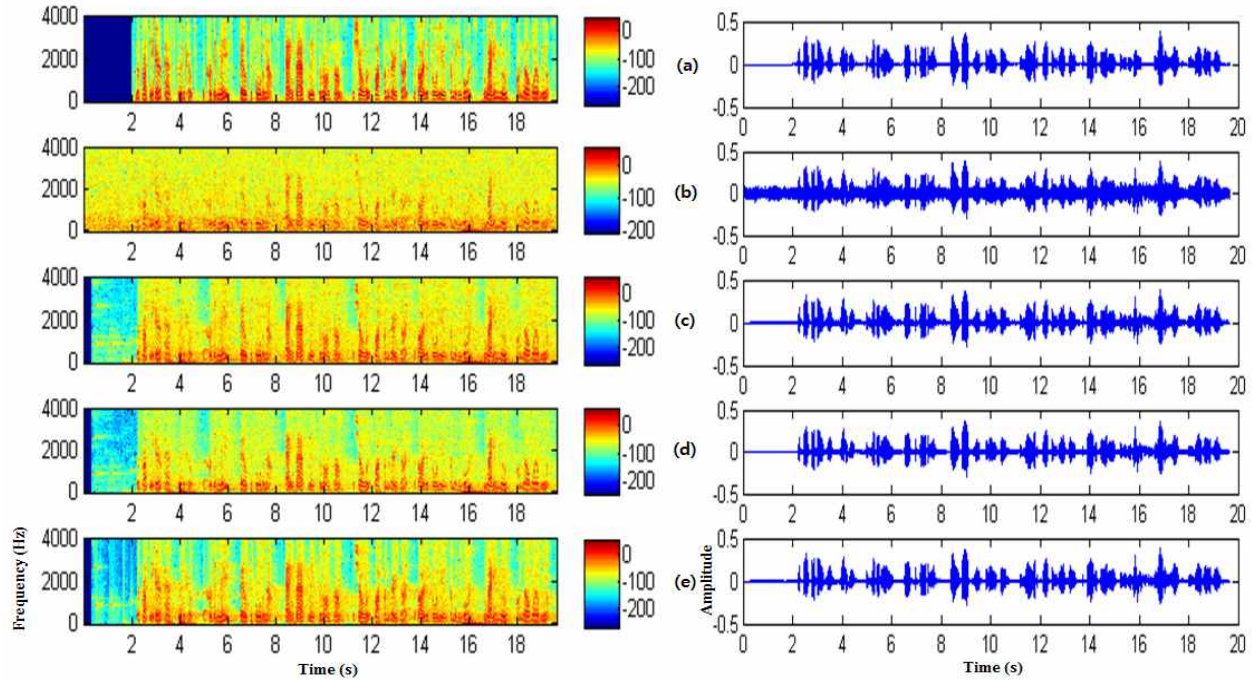

그림 5. 스펙트로그램 및 파형 결과 : 배블 잡음 (a) 깨끗한 음성, (b) 잡음 섞인 입력신호 (c) MVDR결과 (d) 다채널 위너필터 결과 (e) 제안한 최적 변형 다채널 위너필터 결과

를 나타낸다. 두 경우 모두 음성 전력스펙트럼 밀도와 비교해 보았을 때 밀도가 큰 부분에서 다채널 기반 음성존재확률이 1에 가깝고 밀도가 작은 부분에서 다채널 기반 음성 존재 확률이 낮은 것을 확인할 수 있다. 또한, 이론적인 출력 SINR(모의신호를 통해 실험하였기 때문에 섞은 잡음을 알 수 있고 이를 통해 이론적인 출력 SINR을 구할 수 있다)과 비교해 보았을 때 추정된 출력 SINR이 상당히 정확하게 추정되는 것을 확인할 수 있다. 이론과 추정된 출력 SINR의 약간의 차이는 음성과 잡음이 합께 존재하는 상황에서 잡음의 악영향에서 비롯된 것으로 보인다.

잡음 제거 성능을 비교하기 위해, 기존 방식과 제안한 방법으로 잡음제거 후 복원된 음성신호의 파형과 스펙트로그램을 각각 그림 4와 5에 도시하였다. 그림 4와 5의 (c)와 (d)는 각각 MVDR (매개변수 내장형 위너필터에서 매개변수(β)가 0인 경우)과 다채널 위너 필터 (매개변수 내장형 위너필터에서 매개변수(β)가 1인 경우)의 잡음제거 결과를 나타낸다. 다채널 위너필터가 MVDR 대비 잡음을 많이 제거 하지만 음성도 많이 제거하는 것을 확인할 수 있다. 이러한 경향은 그림 5에서 더욱 뚜렷한데 그 이유는 "babble" 이 "F-16" 보다 비정상성이 크기 때문으로 생각된다. 정리하면, 그림 4와 5를 통해서 MVDR은 적은 음성왜곡이 있지만 잔여잡음이 많았고, 다채널 위너필터는 잔여잡음을 줄였으나 음성왜곡이 커졌다. 그러나 제안한 최적 변형 다채널 위너필터의 경우 다채널 위너필터 보다 낮은 음성왜곡을 만족하면서도 잡음을 더 많이 제거할 수 있었다. 그 이유는 음성존재확률을 매개변수 내장형 위너필터의 제어 매개변수로 활용할 수 있었기 때문이다. 제안한 방법의 철학은 해당 주파수 빈에서 음성존재확률이 낮으면 큰 매개변수를 할당하여 잡음제거도를 높이고 반면에 음성존재확률이 높으면 작은 매개변수를 할당하여 잡음제거도를 낮추어 음성왜곡을 줄이는 것이다. 전체 프로세스는 다채널 기반 음성존재확률 값에 따라 자동적으로 제어되는 것이 장점이다.

파형 결과에서도 알 수 있듯이 제안한 방법의 음성구간은 MVDR의 결과와 비슷하고 잡음구간은 다채널 위너필터의 결과와 비슷하다. 추가적으로, 객관적 성능평가를 위해, 기존 방법과 제안한 방법으로 잡음제거 된 음성샘플들의 출력 SINR (oSINR), 잡음제거도 (noise reduction factor, NRF), 음성왜곡도 (signal distortion index, SDI)를 측정하였다 [4-8, 17]. 10세트의 음성샘플 측정값을 평균하여 표 1과 2에 각각 정리하였다. 표 1에서 잡음이 babble일 때, MWF를 기준으로 MVDR의 결과를 비교해 보면, MVDR은 NRF가 MWF보다 약 1.5 dB 낮지만 음성왜곡도는 약 0.6 dB 낮다. MWF 결과 잔여잡음의 양은 줄었지만 음성왜곡도가 높아진 결과로 음성강화에서 흔히 발생하는 트레이드오프 관계에 있음을 알 수 있다. 반면에 OM-MWF는 MWF대비 NRF가 약 0.2 dB 증가, SDI가 약 0.1 dB 감소하였다. 이처럼 NRF와 SDI를 동시에 향상시키는 것은 기존의 필터의 트레이드오프 관계를 벗어나는 결과로서 의미가 있다 할 수 있다. F-16의 경우 OM-MWF가 MWF대비 SDI를 약 0.1 dB 줄이면서 NRF를 약 0.5 dB 향상시켰다. 표 2의 경우도 표1과 비슷한 성능향상 경향을 보였고, 그림 4와 5의 결과와 일관되게, 잡음의 종류와 반향시간과 무관하게 모든 경우에서 제안한 방법이 낮은 음성왜곡도를 만족하면서 높은 출력 SINR과 잡음제거도를 보여주었다.

표 1. 객관적 성능평가(입력 SINR 3.80 dB, 단위 dB)

Filter		MVDR		MWF		OM-MWF	
Interf. Type		babble	F-16	babble	F-16	babble	F-16
T_{60} = 0 ms	oSINR	12.03	12.27	13.39	14.26	13.65	14.92
	NRF	8.32	8.54	9.85	10.70	9.98	11.23
	SDI	-17.31	-17.70	-16.69	-17.15	-16.77	-17.25
T_{60} = 210 ms	oSINR	12.34	12.86	13.89	14.72	14.15	15.44
	NRF	8.86	9.31	10.62	11.35	10.71	11.92
	SDI	-15.21	-15.88	-14.51	-15.38	-14.82	-15.67

표 2. 객관적 성능평가(입력 SINR 8.81 dB, 단위 dB)

필터종류		MVDR		MWF		OM-MWF	
잡음종류		babble	F-16	babble	F-16	babble	F-16
T_{60} = 0 ms	oSINR	15.05	15.27	16.41	16.93	16.73	17.38
	NRF	6.31	6.52	7.80	8.29	8.01	8.65
	SDI	-21.42	-21.26	-20.16	-20.09	-20.44	-20.28
T_{60} = 210 ms	oSINR	15.29	15.61	16.76	17.27	17.09	17.68
	NRF	6.73	7.00	8.32	8.76	8.54	9.09
	SDI	-18.02	-18.27	-17.35	-17.69	-17.75	-18.06

V. 결 론

본 논문은 음성존재확률을 이용하여 매개변수 내장형 위너필터의 이득을 최적으로 변형시키는 방법에 대해서 제안하였다. 제안한 방법을 분석한 결과 매개변수 내장형 위너필터의 매개변수를 음성존재확률에 따라 자동으로 조절할 수 있는 장점이 있었고, 모의실험을 통해 제안한 방법이 기존 방법 대비 babble 잡음의 경우 SDI가 약 0.1 dB 감소하면서 NRF가 약 0.2 dB 향상시켰고, F-16의 경우 SDI를 약 0.1dB 줄이면서 NRF를 약 0.5 dB 향상시키는 결과 통해 음성강화에서 필연적으로 발생하는 음성왜곡도와 잔여잡음의 양 간에 존재하는 트레이드오프에서 벗어나 두 가지 동시에 향상 시킬 수 있음을 증명하였다.

REFERENCES

[1] G. Deepak, J.W. Lee, "Comparison of Two Methods for Stationary Incident Detection Based on Background Image," *스마트미디어저널*, 제1권, 제3호, 48-55쪽, 2012년 9월

[2] 이유라, 김수형, 김영철, 나인섭, "심층 학습 모델을 이용한 EPS 동작 신호의 인식," *스마트미디어저널*, 제5권, 제3호, 35-41쪽, 2016년 9월

[3] P.C. Loizou, *Speech Enhancement: Theory and Practice.* Boca Raton, FL: CRC, pp. 291-394, 2007.

[4] J. Benety, *Microphone Array Signal Processing. Heidelberg*, Berlin: Springer-Verlag, pp. 127-214, 2007.

[5] M. Souden, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 2, pp. 260-276, 2010.

[6] N.S. Kim, J.H. Chang, "Spectral enhancement based on global soft decision," *IEEE Signal Process. Lett.*, vol. 7, no. 6, pp. 108-110, 2000.

[7] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Process. Lett.*, vol. 9, no. 4, pp. 113-116, 2002.

[8] M. Souden, "An integrated solution for online multichannel noise tracking and reduction," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2159-2169, 2011.

[9] M. Souden, "Gaussian model-based multichannel speech presence probability," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 5, pp. 1072-1077, 2010.

[10] IEEE Subcommittee, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. AE-17, no. 3, pp. 225-246, 1969.

[11] A.P. Varga, "The Noisex-92 study on the effect of additive noise on automatic speech recognition," *Tech. Rep. DRA Speech Research Unit*, 1992.

[12] J. Allen, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943-950, 1979.

[13] E. Lehmann, "Prediction of energy decay in room impulse responses simulated with an image-source model," *J. Acoust. Soc. Amer.*, vol. 123, pp. 269-277, 2008.

[14] J.J. Shynk, "Frequency-domain and multirate adaptive filtering," *IEEE Signal Process. Mag.*, vol. 9, no.1, pp. 15-37, 1992.

[15] S. Gannot, "Signal enhancement using beamforming and nonstationarity with application to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614-1626, 2001.

[16] S. Affes, "A signal subspace tracking algorithm for microphone array processing of speech," *IEEE Trans. Speech, Audio Process.*, vol. 5, pp. 425-437, 1997.

[17] J. Chen, "New insights into the noise reduction Wiener filter," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1218-1234, 2006.

저 자 소 개

정상배(정회원)

2002년 한국정보통신대학교 공학부 박사 졸업
2002년 - 2006년 삼성종합기술원 책임연구원
2006년 - 2009년 한국과학기술원 디지털미디어랩 연구조교수
2009년 - 현재 경상대학교 전자공학과 부교수

<주관심분야: 음성인식, 음성/오디오 부호화>

김영일(정회원)

1979년 경북대학교 전자공학과 학사 졸업
1981년 연세대학교 전자공학과 석사 졸업
1985년 연세대학교 전자공학과 박사졸업
1987년 - 현재 경상대학교 전자공학과 교수

<주관심분야: 음향공학, 음성신호처리>