MMSE-STSA 추정치에 기반한 후처리를 갖는 마이크로폰 배열을 이용한 음성 개선

권 홍 석, 손 종 목, 배 건 성
경북대학교 전자·전기공학부

Speech Enhancement Using Microphone Array with MMSE-STSA Estimator Based Post-Processing

Hong Seok Kwon, Jong Mok Son, Keun Sung Bae
School of Electronic and Electrical Engineering, Kyungpook National Unversity
E-mail : hskwon@mir.knu.ac.kr

Abstract

In this paper, a speech enhancement system using microphone array with MMSE-STSA (Minimum Mean Square Error-Short Time Spectral Amplitude) estimator based post-processing is proposed. Speech enhancement is first carried out by conventional delay-and-sum beamforming (DSB). A new MMSE-STSA estimator is then obtained by refining MMSE-STSA estimators from each microphone, which is applied to the output of conventional DSB to obtain additional speech enhancement. Computer simulation for white and pink noises show that the proposed system is superior to other approaches.

I. 서론

음성개선(speech enhancement)은 잡음환경에서 음성인식이나 음성통신의 우수한 성능을 얻기 위하여 전처리과정으로서 사용되어 왔다. 일반적으로 음성개선은 마이크로폰의 개수에 따라 단일채널 기법과 다채널 기법으로 구분된다. 단일채널 기법에는 스펙트럼 차감법[1], Wiener filtering[1], MMSE-STSA 추정치[2] 등을 이용하여 널리 사용되어져 왔다. 그러나 대부분의 단일채널 기법은 잡음전력 추정에 기반하므로, 입력 SNR (Signal-to-Noise Ratio)이 낮은 경우에는 부정확한 잡음추정으로 인하여 성능이 급격히 저하된다. 다채널 기법은 지연합빔형성 (DSB: Delay-and-Sum Beamforming)[3], 적응빔형성(adaptive beamforming), 후처리를 갖는 마이크로폰 배열을 이용하는 기법[4] 등이 있다. DSB는 간단한 구조를 갖지만 우수한 성능을 얻기 위해서는 많은 마이크로폰을 요구한다. 적응빔형성은 우수한 성능에 비해 많은 계산량이 필요하며 잔향이 존재하는 경우에는 성능이 저하된다. 그리고 후처리를 갖는 마이크로폰 배열은 DSB로 음성개선을 한 다음 후처리 과정을 통해 성능을 더욱더 개선시키는 방법이다.

본 논문에서는 MMSE-STSA 추정치에 기반한 후처리 과정을 갖는 마이크로폰 배열을 제시한다. 우선 DSB를 통하여 일차적으로 음성개선을 하고 각 마이크로폰에서 MMSE-STSA 추정치를 위한 파라미터를 계산한다. 계산된 각각의 파라미터로부터 새로운 MMSE-STSA 추정치를 구하여 DSB의 결과에 적용함으로써 더 우수한 음성개선 성능을 얻도록 한다. 음성신호에 백색잡음과 유색잡음을 첨가하여 수행한 모의실험을 통하여 제시한 방법의 성능을 평가하고 그 우수성을 보인다.

II. MMSE-STSA 추정치에 기반한 후처리를 갖는 마이크로폰 배열

2.1 MMSE-STSA 추정치

$x[n]$과 $d[n]$을 각각 음성과 부가잡음으로 표현한다면, 잡음음성, $y[n]$은 식 (1)과 같이 되며 잡음음성의 k번째 스펙트럼, Y_k는 식 (2)와 (3)처럼 크기와

위상 스펙트럼으로 표현할 수 있다.

$$y[n] = x[n] + d[n] \quad (1)$$

$$Y_k = X_k + D_k \quad (2)$$

$$R_k \exp(j\vartheta_k) = A_k \exp(j\alpha_k) + D_k \quad (3)$$

음성과 잡음의 스펙트럼이 서로 통계적으로 독립적인 가우시안 랜덤변수라고 하면, t번째 프레임의 MMSE-STSA 추정치, $\widehat{A_k}$는 식 (4)와 같다[2]. 여기서 사전 SNR(a priori SNR), $\zeta_k(t)$와 사후 SNR(a posteriori SNR), $\gamma_k(t)$는 식 (5)와 (6)으로 주어진다.

$$\begin{aligned}\widehat{A_k}(t) &= E\{A_k(t)\, Y_k(t)\} \\ &= G_{MMSE}(\zeta_k(t), \gamma_k(t)) \cdot R_k(t) \quad (4)\end{aligned}$$

$$\gamma_k(t) \equiv \frac{R_k^2(t)}{\lambda_{dk}(t)} \quad (5)$$

$$\zeta_k(t) \equiv \frac{\lambda_{xk}(t)}{\lambda_{dk}(t)} = \alpha \frac{\widehat{A_k^2}(t-1)}{\lambda_{dk}(t)} + (1-\alpha) P[\lambda_k(t) - 1] \quad (6)$$

여기서, $\lambda_{dk}(t)$와 $\lambda_{xk}(t)$는 각각 잡음과 음성의 k번째 주파수 성분을 말하며, α는 망각지수이고 $P[\cdot]$는 양의 값을 보장하기 위한 연산자이다. 사후 SNR은 잡음 음성과 추정잡음의 분산으로부터 직접 구하며, 사전 SNR은 "Decision-Directed" 추정방법을 이용하여 구할 수 있다[2]. 이득, G_{MMSE}는 식 (4)와 같이 사전 SNR과 사후 SNR만의 함수로 표현된다. 이렇게 구해진 MMSE-STSA 추정치에 음성존재확률을 도입하면 식 (7)과 같이 수정된 MMSE-STSA 추정치를 구할 수 있다.

$$\begin{aligned}\widehat{A_k}(t) &= P(H_k^1 | Y_k(t)) \cdot G_{MMSE}(\zeta_k(t), \gamma_k(t)) \cdot R_k(t) \\ &= \frac{\Lambda_k(t)}{1+\Lambda_k(t)} \cdot G_{MMSE}(\zeta_k(t), \gamma_k(t)) \cdot R_k(t) \quad (7)\end{aligned}$$

여기서 $P(H_k^1 | Y_k(t))$는 주어진 $Y_k(t)$에 대한 음성존재확률로서, 식 (8)처럼 우도(likelihood ratio)로 정의된다. 식 (8)에서 H_k^1과 H_k^0는 각각 음성의 존재와 부재에 대한 가설(hypothesis)고 $P(H_k^0)$는 k번째 주파수 성분의 사전 음성부재확률이다.

$$\begin{aligned}\Lambda_k(t) &\equiv \frac{1-P(H_k^0)}{P(H_k^0)} \cdot \frac{P(Y_k(t)|H_k^1)}{P(Y_k(t)|H_k^0)} \quad (8) \\ &= \frac{1-P(H_k^0)}{P(H_k^0)} \cdot \frac{1}{1+\zeta_k(t)} \exp\left[\frac{\gamma_k(t)\zeta_k(t)}{1+\zeta_k(t)}\right]\end{aligned}$$

본 논문에서는 음성의 검출실패로 인한 음질저하를 줄이기 위하여 HMM에 기반한 hangover 방법도 적용하였다[5]. 그리고 사전 SNR과 사후 SNR을 추정하기 위해 필요한 잡음의 분산은 식 (9)와 (10)처럼 사전 SNR을 이용하는 soft-decision에 기반한 방법을 사용하고 smoothing하여 사용하였다[6]. 식 (10)에서 $G_{nk}(t)$는 최적 Wiener 필터의 주파수 특성으로서, 간단히 식 (11)처럼 사전 SNR로부터 구할 수 있다. 그리고 η는 잡음의 분산을 smoothing하기 위한 망각지수이다.

$$\lambda_{dk}(t) = \eta \lambda_{dk}(t-1) + (1-\eta) E(|D_k(t)|^2 | Y_k(t)) \quad (9)$$

$$\begin{aligned}&E(|D_k(t)|^2 | Y_k(t)) \quad (10) \\ &= \{P(H_k^0 | Y_k(t)) + P(H_k^1 | Y_k(t)) \cdot G_{nk}^2(t)\} \cdot |Y_k(t)|^2\end{aligned}$$

$$G_{nk}(t) = \frac{E(|D_k(t)|^2)}{E(|D_k(t)|^2) + E(|X_k(t)|^2)} = \frac{1}{1+\zeta_k(t)} \quad (11)$$

2.2 MMSE-STSA 추정치에 기반한 후처리

그림 1은 본 논문에서 제시한 MMSE-STSA에 기반한 후처리를 갖는 마이크로폰 배열을 보여주고 있다. 먼저 DSB를 이용하여 일차적으로 음성개선을 수행하며, MMSE-STSA 추정치를 구하기 위해 필요한 파라미터로서 사전 SNR, 사후 SNR, 우도를 각 마이크로폰에서 구한다. 이를 이용하여 새로운 MMSE-STSA 추정치를 구하여 DSB의 출력에 적용함으로써 더욱더 음성을 개선시킨다.

MMSE-STSA 추정치를 구하기 위한 이득은 식 (4)를 보면 알 수 있듯이 사전 SNR과 사후 SNR에 의해서만 결정된다. 즉 정확한 사전 SNR과 사후 SNR에 의해서만 정확한 이득을 계산할 수 있다. 따라서, 새로운 사전 SNR과 사후 SNR을 식 (12)와 (13)과 같이 각 마이크로폰에서 구한 사전 SNR과 사후 SNR의 평균을 취함으로써 얻는다. 식 (12)와 (13)에서 $\gamma_{km}(t)$와 $\zeta_{km}(t)$는 M개의 마이크로폰으로 구성된 마이크로폰 배열의 m번째 마이크로폰에서 구한 사후 SNR과 사전 SNR을 말한다.

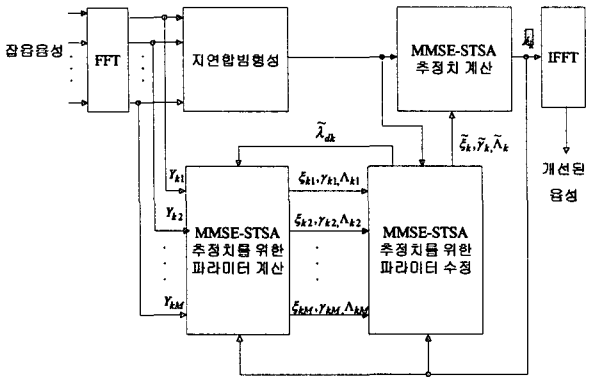

그림 1. MMSE-STSA에 기반한 후처리를 갖는 마이크로폰 배열

$$\widetilde{\gamma_k}(t)=\frac{1}{M}\sum_{m=1}^{M}\gamma_{km}(t) \tag{12}$$

$$\widetilde{\zeta_k}(t)=\frac{1}{M}\sum_{m=1}^{M}\zeta_{km}(t) \tag{13}$$

식 (7)을 보면 알 수 있듯이 음성존재확률을 이득에 곱해서 MMSE-STSA 추정치를 구하게 되므로 음성존재확률은 음성개선의 성능에 직접적인 영향을 미친다. 따라서 우수한 성능을 위해서는 정확한 음성존재확률을 구하는 것이 필요하다. 식 (7)로부터 정확한 음성존재확률은 정확한 우도를 계산함으로써 가능하다. 본 논문에서는 식 (14)를 이용하여 각 마이크로폰으로부터 구해진 우도의 기하평균을 취함으로써 새로운 우도를 구한다. 여기서 $\Lambda_{km}(t)$는 m번째 마이크로폰에서 구한 우도를 말한다.

$$\widetilde{\Lambda_k}(t)=\frac{1-P(H_k^0)}{P(H_k^0)}\cdot\left(\prod_{m=1}^{M}\Lambda_{km}(t)\right)^{\frac{1}{M}} \tag{14}$$

하나의 마이크로폰을 이용하여 MMSE-STSA 추정치를 구할 때, 식 (5), (9), (10)을 보면 알 수 있듯이 사전 SNR과 잡음의 분산은 서로 의존성을 갖는다. 따라서 본 논문에서는 잡음의 분산과 사전 SNR의 의존성을 피하기 위하여 식 (11)의 최적 Wiener 필터를 구할 때, 식 (15)와 같이 자기전력스펙트럼(auto power spectrum)과 상호전력스펙트럼(cross power spectrum)을 이용하였다[4]. 각 마이크로폰의 잡음 스펙트럼이 서로 상관성이 없고 음성의 스펙트럼과 상관성이 없다고 가정하면, 식 (15)의 자기전력스펙트럼과 상호전력스펙트럼은 식 (16)으로 구할 수 있다. 이는 잡음의 분산과 사전 SNR의 의존성을 제거하게 되며, 최종적으로 식 (17)처럼 각 마이크로폰의 잡음의 분산을 평균함으로써 더 정확한 잡음의 분산을 얻는다.

$$G_{nk}(t)=\frac{E(|D_k(t)|^2)}{E(|D_k(t)|^2)+E(|X_k(t)|^2)}=1-\frac{R_k^{xx}(t)}{R_k^{yy}(t)} \tag{15}$$

$$R_k^{yy}(t)=\frac{1}{M}\sum_{m=1}^{M}|Y_{km}(t)|^2 \tag{16}$$

$$R_l^{xx}(t)=Re\left(\frac{2}{M(M-1)}\sum_{m=1}^{M-1}\sum_{l=m+1}^{M}Y_{km}(t)\,Y_{kl}^{*}(t)\right)$$

$$\widehat{\lambda_{dk}}(t)=\frac{1}{M}\sum_{m=1}^{M}\lambda_{dkm}(t) \tag{17}$$

Ⅲ. 실험결과

제안한 방법의 성능을 평가하기 위하여 DSB의 출력, 제안한 방법 그리고 Zelinski 방법[4]의 결과를 평균 출력 SNR과 LAR 왜곡에 대하여 서로 비교하였다. 신호는 4개의 선형 마이크로폰 배열에 수직으로 입사된다고 가정하였으며, DSB에서 입사각도로 조향하기 위하여 100kHz로 샘플링되고 16비트로 양자화한 다음, 이를 10kHz로 다운샘플링하여 음성개선에 사용한다고 가정하였다. 한 프레임은 Hanning 윈도우를 이용하여 256샘플로 구성하고 64샘플씩 이동시켰으며, 256포인트 FFT를 사용하였다. 잡음음성은 3개의 다른 음성에 백색잡음과 유색잡음인 pink 잡음을 더하여 100kHz에서 5dB 간격으로 -15dB에서 15dB를 갖도록 만들었다. 그리고 DSB에서 입사각도로 조향할 때 어떠한 오차도 발생하지 않는다고 가정하였다.

그림 2와 3은 백색잡음과 유색잡음에 대하여 3개의 음성으로 구한 평균 출력 SNR을 나타낸 것이다. 이때 각각의 잡음제거기법에 사용되는 파라미터는 실험적으로 구하였다. 그림 2로부터 백색잡음에서는 제안한 방법이 Zelinski 방법보다 약 1dB에서 2dB 정도 우수함을 알 수 있다. 그림 3의 유색잡음인 경우에는 제안한 방법이 낮은 SNR에서는 Zelinski 방법과 유사하지만 높은 SNR에서는 제안한 방법이 더 높은 출력 SNR을 가졌다. 따라서 백색잡음과 유색잡음에 대해 제안한 방법이 다른 방법에 비해서 우수한 성능을 갖는다고 말할 수 있다. 그림 4와 5는 백색잡음과 유색잡음에 대하여 사람의 청각특성을 잘 반영한다고 알려진 LAR 왜곡을 3개의 음성으로 구한 평균 LAR 왜곡을 나타낸 것이다. 그림으로부터 제안한 방법이 백색잡음과 유색잡음에서 다른 방법에 비해서 낮은 LAR 왜곡을 가짐을 알 수 있다. 그리고 주관적인 청취실험을 통해서 제안한 방법이 잡음이 더 많이 제거되고 musical 잡음도 적음을 알 수 있었다.

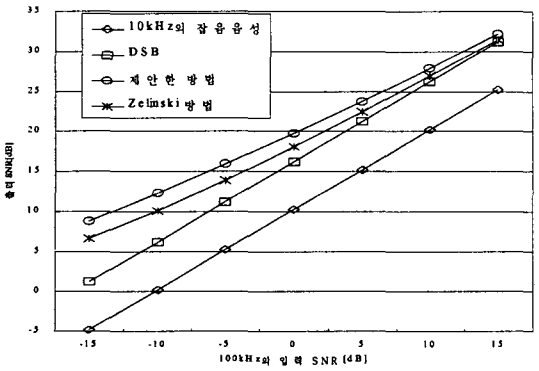

그림 2. 백색잡음에서 평균 출력 SNR

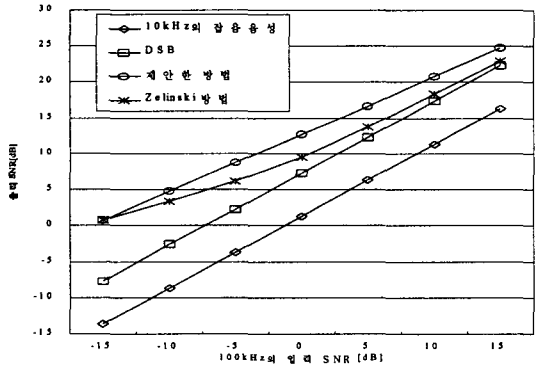

그림 3. 유색잡음에서 평균 출력 SNR

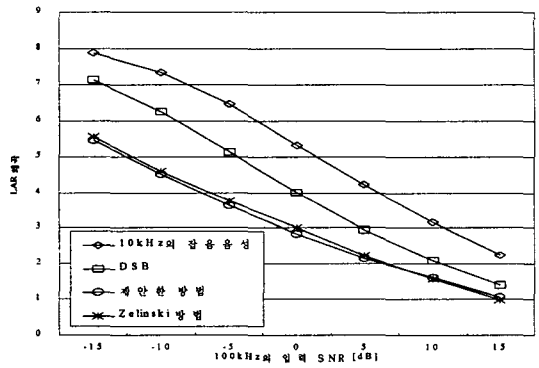

그림 4. 백색잡음에서 평균 LAR 왜곡

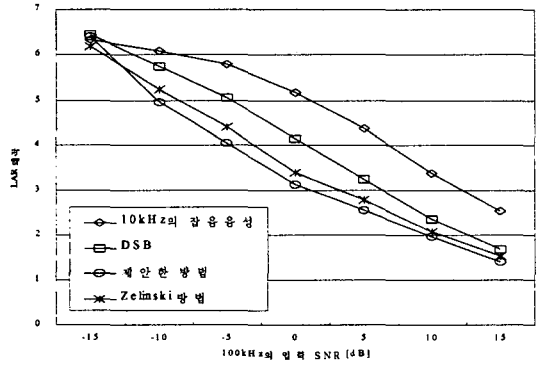

그림 5. 유색잡음에서 평균 LAR 왜곡

Ⅳ. 결론

본 논문에서는 MMSE-STSA 추정치에 기반한 후처리를 갖는 마이크로폰 배열을 이용한 음성개선 방법을 제시하였다. DSB에 의해서 일차적으로 음성개선을 하고, 각 마이크로폰 배열에서 MMSE-STSA 추정치를 위한 변수를 구하여 새로운 MMSE-STSA 추정치를 구하였다. 이를 DSB의 출력에 적용함으로써 더 높은 음성개선 성능을 얻었다. 결과적으로 백색잡음과 유색잡음으로 평가한 평균 출력 SNR과 LAR 왜곡에서 제안한 방법이 다른 방법에 비해서 성능이 우수한 성능을 보였다.

본 연구는 한국과학재단의 목적기초연구(과제번호: R01-1999-00233) 연구비 지원으로 수행 되었습니다.

참고문헌

[1] S.J. Godsill, and P.J.W. Rayner, *Digital Audio Restoration*, Springer, 1998.

[2] Y. Ephraim, and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. on ASSP*, Vol. ASSP-32, No. 6, pp. 1109-1121, Dec. 1984.

[3] B.D. Van Veen, and K.M. Buckley, "Beamforming: a versatile Approach to Spatial Filtering," *IEEE ASSP Magazine*, Vol. 5, No. 2, pp. 4-24, Apr. 1988.

[4] R. Zelinski, "A Microphone Array with Adaptive Post-Filtering for Noise Reduction in Reverberant Rooms," *ICASSP*, pp. 2578-2581, 1988.

[5] 장준혁, 김남수, "행오버를 이용한 Soft Decision 음성향상기법," *한국방송공학회학술대회*, pp. 201-206, 1999.

[6] Y.D. Cho, K. Al-Naimi, and A. Kondoz, "Mixed Decision-Based Noise Adaptation for Speech Enhancement," *Electronics letters*, Vol. 37, No. 8, pp. 540-542, 2001.