Regular paper
Journal of the KIECS. pp. 465-470, vol. 16, no. 3, Jun. 30. 2021, t. 107, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2021.16.3.465

음성 통계 모형에 따른 음성 왜곡량 감소를 위한 비선형 음성강조법

최재승*

Nonlinear Speech Enhancement Method for Reducing the Amount of Speech Distortion According to Speech Statistics Model

Jae-Seung Choi*

요 약

잡음이 존재하는 실제 환경에서 음성인식을 실시하는 경우에 음성인식의 성능 열화 및 음성의 품질이 저화되지 않는 강건한 음성인식 기술이 필요하다. 이러한 음성인식 기술을 개발함으로써 사람의 음성 스펙트럼과 유사한 잡음 환경에서도 안정되고 높은 음성인식률이 실현되는 어플리케이션이 요구된다. 따라서 본 논문에서는 최소 평균 제곱의 오차를 기반으로 한 단시간 스펙트럼 진폭 방법인 MMSA-STSA 추정 알고리즘에 기초한 잡음억압을 처리하는 음성강조 알고리즘을 제안한다. 이 알고리즘은 단일 채널 입력에 기초한 효과적인 비선형 음성강조 알고리즘이며, 높은 잡음억제 성능을 가지고 있으며 음성의 통계적인 모델에 기초하여 음성의 왜곡량을 줄이는 기법이다. 본 실험에서는 MMSA-STSA 추정 알고리즘의 유효성을 확인하기 위하여 입력 음성파형과 출력 음성파형을 비교하여 제안한 알고리즘의 효과를 확인한다.

ABSTRACT

A robust speech recognition technology is required that does not degrade the performance of speech recognition and the quality of the speech when speech recognition is performed in an actual environment of the speech mixed with noise. With the development of such speech recognition technology, it is necessary to develop an application that achieves stable and high speech recognition rate even in a noisy environment similar to the human speech spectrum. Therefore, this paper proposes a speech enhancement algorithm that processes a noise suppression based on the MMSA-STSA estimation algorithm, which is a short-time spectral amplitude method based on the error of the least mean square. This algorithm is an effective nonlinear speech enhancement algorithm based on a single channel input and has high noise suppression performance. Moreover this algorithm is a technique that reduces the amount of distortion of the speech based on the statistical model of the speech. In this experiment, in order to verify the effectiveness of the MMSA-STSA estimation algorithm, the effectiveness of the proposed algorithm is verified by comparing the input speech waveform and the output speech waveform.

키워드

Speech Recognition, Speech Enhancement Algorithm, Noise Suppression, MMSA-STSA Estimation, Statistics Model
음성 인식, 음성 강조 알고리즘, 잡음 억압, MMSA-STSA 추정, 통계적 모델

* 신라대학교 스마트전기전자공학부
• 접 수 일 : 2021. 04. 28
• 수정완료일 : 2021. 05. 23
• 게재확정일 : 2021. 06. 17
• Received : Apr. 28, 2021, Revised : May. 23, 2021, Accepted : Jun. 17, 2021
• Corresponding Author : Jae-Seung Choi
Division of Smart Electrical and Electronic Engineering, Silla University
Email : jschoi@silla.ac.kr

Ⅰ. 서 론

음성신호는 사람이 가장 자연스럽게 사용하는데 편리한 통신수단 중의 하나이며, 근년에는 휴대용 이동통신전화, 음성인식대화 로봇 및 TV 통신회화 시스템이라고 하는 음성통신에 관한 시스템의 이용이 확대되고 있다. 또한 디지털 보청기, 로봇대화 및 Car Navigation 시스템에 대한 음성인식 등 음성을 이용한 다양한 어플리케이션이 연구되고 있다. 이러한 어플리케이션에서는 실제 잡음이 존재하는 환경 하에서도 목적으로 하는 음성의 청각 품질 및 음성인식 성능 등이 저화 되는 것이 없이 강건하게 동작하는 것을 전제로 하고 있다[1-2].

현재, 이러한 음성시스템을 실현하기 위해서는 이용자가 입력되는 마이크로폰으로부터 떨어질 수 없는 상황이기 때문에 주변 환경의 소음에 강한 접화형 마이크로폰 및 헤드셋 마이크로폰을 이용할 필요가 있다. 이와 더불어 실제 환경에 있어서는 목적음성 이외의 잡음이 많이 존재하여 이러한 잡음이 마이크로폰에 동시에 혼입하여 들어오기 때문에 목적으로 하는 음성의 품질을 열화시키는 문제를 발생시킨다. 따라서 핸즈프리 음성처리시스템의 실현에 있어서 잡음을 포함한 관측된 신호로부터 목적으로 하는 음성만을 높은 정밀도로 추출하는 음성강조법의 실현이 필요하게 된다[3-4].

인간은 다양한 음성이 혼재하는 중에서도 목적으로 하는 음성을 청취하여 분류할 수 있는 Cocktail Party효과는 음원분리 및 잡음억압에 관한 기술 분야에서 연구가 활발히 수행되고 있으며 다양한 수법이 제안되고 있다[5]. 이러한 기술을 이용함으로서 잡음환경 하에서도 목적으로 하는 음성신호만을 추출하는 것이 가능하게 되고, 고음질인 목적 음성의 청취 및 높은 음성인식 성능을 실현 가능할 수 있게 한다. 따라서 이러한 기술을 실현함으로써 잡음환경에 크게 상관없이 안정되게 동작되는 강건한 음성인식 어플리케이션의 개발이 가능할 수 있게 된다.

잡음억압을 목적으로 한 음성강조수법 중에는 Boll의 스펙트럼 차감법(SS: Spectral Subtraction)[6]이 제안되어 있다. SS는 수신신호의 진폭 스펙트럼으로부터 잡음의 진폭스펙트럼을 추정하여 진폭 스펙트럼의 평균치를 차감함으로써 잡음억압을 실현하는 기법이다. 백색잡음과 같은 정상잡음만을 대상으로 하고 있으며 비정상적인 잡음은 대상으로 하고 있지 않다. 또한 잡음의 추정오차 등으로부터 발생하는 음악성 잡음이 문제가 될 수 있지만 처리가 대단히 간결하기 때문에 현재까지도 여러 분야에서 다양한 개량법이 제안되어 개발되고 있다[7].

통계적 신호처리에 기초한 수법으로는 Ephraim과 Malah에 의해 제안된 최소 평균 제곱의 오차를 기반으로 한 단시간 스펙트럼 진폭(MMSA-STSA: Minimum Mean-Square Error Short-Time Spectral Amplitude)[8] 기법이 있으며, 이 MMSA-STSA는 단시간에서 추정된 진폭 스펙트럼의 평균 제곱의 오차를 최소로 한다. MMSA-STSA에서는 잡음환경 하에서의 음성강조를 가능하게 하지만 비상정상잡음에 대해서는 상당히 약하다는 면도 있다. Lim과 Oppenheim에 의해 제안된 수법에서는 Wiener가 제안한 Wiener filter를 음성에 적용하고 있다[9]. 이 수법에서는 프레임 간에서 추정한 잡음의 진폭 스펙트럼에 대해서 평균 제곱의 오차를 계산하여 최적인 이득 함수를 합성한다. 이러한 수법들은 목적신호가 서로 상관이 없다는 점을 가정하고 있기 때문에 잔향환경 하에서는 적용하기가 어렵다는 어려움이 있다.

본 논문에서는 MMSA-STSA 추정 알고리즘에 기초하여 잡음억압을 처리하는 음성강조 알고리즘을 제안한다. 특히 음성강조법 중에서 중요한 문제의 하나로서 열거되고 있는 단일 채널 입력에 기초한 효과적인 비선형 음성강조법인 MMSE-STSA 추정 알고리즘을 제안한다. 제안하는 단일 채널 비선형 음성강조 알고리즘인 MMSE-STSA 측정기는 높은 잡음억제 성능을 가지고 있고 음성의 통계적인 신호처리의 사전 모델에 기초한 것이며, 스펙트럼 감산법 및 위너 필터링과 비교하여 음악잡음 발생량, 음성 왜곡량 등이 적은 음성강조법이기 때문에 제안 수법으로부터 음성품질의 개선이 기대된다. 더욱이 MMSE-STSA 측정기에 기초한 잡음 음성강조법의 유효성을 확인하기 위하여 출력 음성파형으로 음성강조법의 실험을 실시한다.

전체 논문의 구성은 2장에서 기존의 잡음 억압에 대한 연구를 기술하였다. 3장에서는 제안하는 알고리즘에 대한 개요 및 실험 내용을 기술하였으며, 4장에서는 결론을 나타내었다.

Ⅱ. 종래의 잡음 억압 연구

근래에 음성강조는 음성 특징 추출을 기본으로 하여 여러 분야에서 신호처리 기술을 이용하여 발전시켜 왔으

며, 이러한 음성강조 기법을 음성의 어플리케이션 및 디지털 보청기 시스템에 도입하는 것에 의해서 이러한 성능을 높이는 것이 가능하였다.

음성신호 데이터의 품질 및 성능의 개량을 목적으로 한 잡음 억제 신호처리 기법의 연구가 제안되고 있으며 여러 분야에서 이용되고 있다. 정상 잡음을 가법성 잡음으로 가정하면 잡음이 포함된 관측신호 $y(t)$는 음성신호 $v(t)$와 잡음신호 $n(t)$를 사용하여 식 (1)과 같이 정의할 수 있다.

$$y(t) = v(t) + n(t) \tag{1}$$

식 (1)은 단시간 푸리에 변환한 주파수 영역을 표현하고 있다. 제 j프레임에 있어서 프레임의 선두로부터 m번째 주파수의 관측신호의 복소 스펙트럼 $Y(j,m)$에 대해서 식(2)와 같이 관계식이 구해진다.

$$Y(j,m) = V(j,m) + N(j,m) \tag{2}$$

음악잡음의 발생은 음질에 부여되는 영향의 크기로부터 비선형잡음 억압처리에 대하여 중대한 문제가 되며, 음악잡음의 발생을 억제하는 수법이 필요하게 된다. 비교적 음악잡음의 발생량이 적은 음성강조법으로서 위너 필터링(Wiener Filtering)[10]이 제안되고 있다. 이 수법은 스펙트럼 감산법과 비교하여 음악잡음의 발생량이 적은 장점이 있지만 사람이 청취해야 할 용도로 사용하기에는 곤란한 점도 가지고 있다. 이와 같이 단일 채널 비선형 음성강조법에 있어서 음악잡음이 완전히 발생하지 않는 잡음억압수법을 찾는 것이 지금까지 쉽지 않다고 할 수 있다.

위너 필터링에 의한 잡음억제 처리는 관측신호의 스펙트럼 $Y(j,m)$에 이득 함수 $G(j,m)$을 적용하여, 음성신호의 스펙트럼의 추정치 $\hat{V}(j,m)$을 식 (3)과 같이 구함으로서 잡음을 억제하는 수법이다.

$$\hat{V}(j,m) = G(j,m)\,Y(j,m) \tag{3}$$

여기에서 이득 함수 $G(j,m)$는 필터이며 신호처리의 수법에 의해서 다르게 적용할 수 있다. 위너필터링법의 이득 함수 $G(j,m)$는 식 (4)와 같이 구해진다.

$$G(j,m) = \frac{\Gamma(j,m)}{1+\Gamma(j,m)} \tag{4}$$

여기에서 $\Gamma(l,k)$는 사전 신호대잡음비(Signal-to-Noise Ratio, SNR) 이라고 부르며 식 (5)와 같이 SNR의 추정치로 표현할 수 있다.

$$\Gamma(j,m) = \frac{|\hat{V}(j,m)|^2}{|\hat{N}(j,m)|^2} \tag{5}$$

단일 채널입력에 기초한 대표적인 음성강조법으로서의 SS는 적은 감산량으로 높은 잡음억제성능을 달성하는 것이 가능하기 때문에 오래전부터 연구된 기술이다. 그러나 이 SS는 주로 비선형처리에 의해서 청각적으로 상당히 불유쾌한 음악잡음으로 알려진 왜곡이 발생하는 문제를 가지고 있다.

최소 평균 제곱의 오차를 기반으로 한 단시간 스펙트럼 진폭 방법인 MMSA-STSA 추정법은 관측신호로부터 계산된 사전 및 사후 SNR을 이용하여 실제의 음성진폭 스펙트럼과 추정된 음성신호의 진폭 스펙트럼과의 평균 자승 오차를 최소화함으로서 스펙트럼 이득을 구한다. 구해진 스펙트럼 이득을 관측신호의 스펙트럼에 승산함으로서 목적의 음성신호의 추정치를 구하는 방법이다[8].

독립성분분석(ICA: Independent Component Analysis)에 기초한 블라인드 음원분리 기법이 제안되어 있다[3, 11]. ICA는 복수의 음원이 통계적으로 독립되어 있다는 가정을 기초로 하여, 수신신호를 복수의 가법적 성분으로 분리하는 가산수법이다. 또한 복수의 마이크로폰 소자를 사용하여 ICA를 실시하는 수법이 제안되어 있다. 이 수법은 분리 가능한 잡음음원의 수가 마이크로폰 소자수와 동등 혹은 그 이하로 제한되어 있기 때문에, 이로 인하여 잡음원의 수가 미지인 실험환경에의 적용을 고려하는 경우 시스템의 규모가 대규모로 될 가능성이 있을 수 있다. 잡음잔향환경 하에서 복수의 마이크 소자를 이용한 음성강조수법은 음성시스템의 인식률을 향상시키는 것이 가능하지만 고역의 주파수 대역에 잡음성분이 남을 수 있다. 이 때문에 사람이 청취하는 것을 상정하는 어플리케이션에의 적용에는 의문이 남을 수 있다.

III. 제안하는 알고리즘 및 실험

MMSE-STSA 추정 알고리즘은 최소 평균의 제곱 오차를 기반으로 한 단시간 스펙트럼 진폭에 의하여 추정하는 기법이다. 이 MMSE-STSA는 음성 및 잡음 스펙트럼 구성요소를 통계적으로 독립적인 가우스 랜덤변수로 모델링하여 도출 할 수 있다. MMSE-STSA 추정기의 중요한 특성은 향상된 음성 신호에서 음악적 잡음을 줄일 수 있다는 것이다[12].

잡음에 대해서 강건한 핸즈프리 음성처리시스템을 실현하기 위해서 여러 음성강조법이 제안되고 있다. 이러한 음성강조법은 잡음환경에서의 기법이기 때문에 잡음억압을 기초로 하고 있다. 기본적인 시간신호는 $y(t)=v(t)+n(t)$의 관계식을 사용한다. 이 시간신호에 대한 주파수영역의 스펙트럼은 $Y(j,m)= V(j,m)+N(j,m)$로 표현할 수 있다. 일반적으로 추정 입력 스펙트럼은 식 (6)과 같이 표현할 수 있다.

$$\hat{V}(j,m) = G(j,m)\,Y(j,m) \qquad (6)$$

여기에서 $G(j,m)$은 잡음이 포함된 입력혼합신호와 잡음신호의 스펙트럼의 차를 잡음입력혼합신호의 스펙트럼으로 나눈 이상적인 이득함수를 나타낸다.

m 번째의 푸리에 확장 계수에 대한 진폭 추정기 $\hat{A}_m$은 로그 전력 스펙트럼의 평균 제곱 오차를 최소화하여 구한다. 희망하는 진폭 추정기 $\hat{A}_m$는 식 (7)과 같이 구할 수 있다.

$$\hat{A}_m \triangleq G(\zeta_m,\gamma_m)R_m \qquad (7)$$

여기에서 Ephraim과 Malah에서 유도 된 이득 함수 $G(\zeta_m,\gamma_m)$는 로그 스펙트럼의 평균 제곱 오차 추정을 최소화하기 위해 구성된다. 마지막으로, m 번째 음성 스펙트럼 성분의 추정량 $\widehat{V_m}$은 식 (7)을 사용하여 식 (8)과 같이 구해진다.

$$\widehat{V_m} \triangleq G(\zeta_m,\gamma_m)\,Y_m \qquad (8)$$

본 실험에서는 제안한 알고리즘을 분석하고 실험결과를 테스트하기 위하여, 잡음이 혼합된 음성 데이터베이스인 Aurora2 Database를 사용한다. 이 Aurora2 Database[13]는 여러 종류의 잡음을 사용하여 다양한 SNR level(20dB, 15dB, 10dB, 5dB, 0dB, -5dB)에 대해서 잡음이 혼합된 음성신호가 준비되어 있다. Aurora2 Database는 깨끗한 녹음환경에서 녹음된 숫자 데이터베이스를 사용하고 각 잡음조건에 따라서 숫자를 포함한 잡음음성 파일로 구성되어 있다. Aurora2 Database는 8kHz 표본화율에 샘플 당 16bits로 양자화된 음성신호로 구성된다.

본 논문의 MMSE-STSA의 신호 및 잡음 전력 추정에 따른 음성개선 분석 실험에는 Aurora2 Database의 Clean speech 신호를 사용하며, 잡음음성으로는 babble noise를 혼합한 잡음음성을 사용하여 실험을 수행한다. 본 실험에서는 20dB, 15dB, 10dB의 SNR을 사용하여 입력 음성신호를 구성하여 실험을 실시하였다.

본 실험 결과로서 본 논문에서 제안한 MMSE-STSA 알고리즘을 사용하여 음성을 강조한 파형을 그림에 나타낸다. 그림 1과 그림 2는 본 실험에서 사용한 Clean speech signal과 Babble noise를 각각 나타내고 있다. 그림 3과 그림 4는 Babble noise가 Clean speech signal에 부가된 Noisy speech signal을 나타내고 있으며, 각각 SNR=20dB, SNR=10dB의 잡음혼합신호에 해당한다. 그림 5와 그림 6은 MMSE-STSA 알고리즘에 의하여 강조된 출력 음성신호(Enhanced speech signal)를 각각 나타내고 있다. 그림 5와 그림 6의 Enhanced speech signal에서 알 수 있듯이 잡음이 증가함에 따라서 잡음제거의 효과가 줄어들긴 하지만 전반적으로 본 논문에서 제안한 MMSE-STSA 추정 알고리즘이 유효하다는 것을 본 실험을 통해서 확인할 수 있었다. 따라서 사람의 음성 스펙트럼과 유사한 특징을 가지고 있는 Babble잡음에 대해서 SNR 레벨이 낮은 SNR=10dB 경우에도 본 논문에서 제안한 알고리즘이 성능이 음성의 특징을 왜곡시키지 않고 효과가 있다는 것을 나타내고 있다.

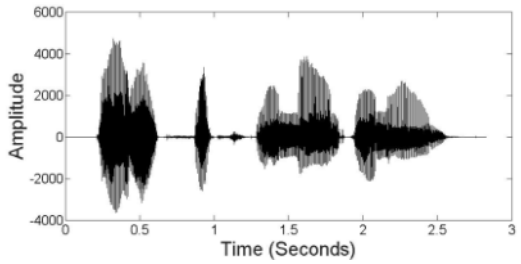

그림 1. 깨끗한 입력 음성신호
Fig. 1 Input clean speech signal

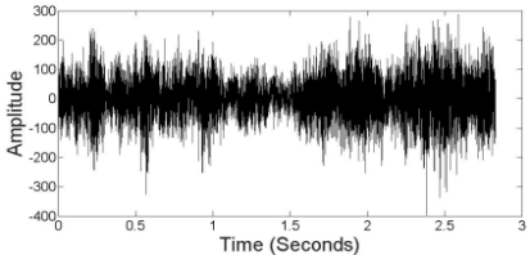

그림 2. 바블 잡음신호
Fig. 2 Babble noise signal

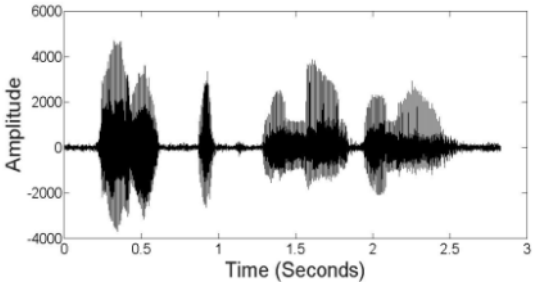

그림 3. 바블잡음이 혼합된 입력 음성신호(SNR=20dB)
Fig. 3 Input noisy speech signal with babble noise (SNR=20dB)

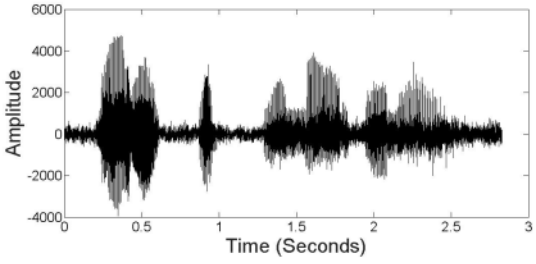

그림 4. 바블잡음이 혼합된 입력 음성신호(SNR=10dB)
Fig. 4 Input noisy speech signal with babble noise (SNR=10dB)

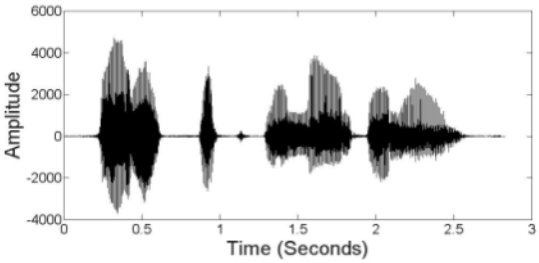

그림 5. 강조된 출력 음성신호(SNR=20dB)
Fig. 5 Enhanced output speech signal (SNR=20dB)

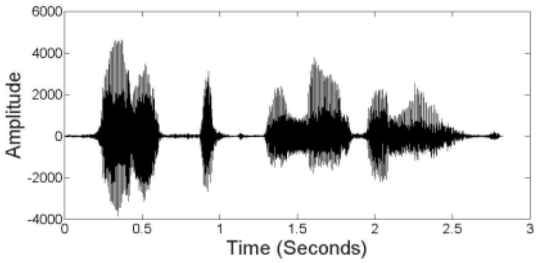

그림 6. 강조된 출력 음성신호(SNR=10dB)
Fig. 6 Enhanced output speech signal (SNR=10dB)

Ⅳ. 결 론

본 논문에서는 잡음환경에 강인한 MMSE-STSA에 의한 음성개선 기법을 사용하여 잡음억압을 처리하는 음성강조 알고리즘을 제안하였다. 본 논문에서 제안한 MMSE-STSA 추정에 의한 음성강조 알고리즘은 높은 잡음억제 성능을 가지고 있고 음성의 통계적인 사전 모델에 기초한 것이며 원음성의 왜곡량 등이 적은 음성강조법이다.

제안한 방법들의 성능을 평가하기 위해 Aurora2의 음성 데이터베이스 및 babble noise를 사용하여 성능향상을 확인하였다. 따라서 본 논문에서 제안한 MMSE-STSA 추정 알고리즘이 음성개선에 효과적임을 볼 수 있었으며, 더욱이 MMSE-STSA 추정 알고리즘에 기초한 잡음 음성강조법의 유효성을 확인할 수 있었다.

References

[1] J. H. L. Hansen and M. A. Clements, "Constrained Iterative Speech Enhancement with Application to Speech Recognition," *IEEE*

Transactions on Signal Processing, vol. 39, no. 4, Apr. 1991, pp. 795-805.

[2] H. Lee, "Acoustic Feedback and Noise Cancellation of Hearing Aids by Deep Learning Algorithm," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 14, no. 6, Dec. 2019, pp. 1249-1256.

[3] J. Choi, "Independent Component Analysis based on Frequency Domain Approach Model for Speech Source Signal Extraction," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 15, no. 5, Oct. 2020, pp. 807-812.

[4] C. Lee, "Dimensionality Reduction in Speech Recognition by Principal Component Analysis," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 8, no. 9, Sept. 2013, pp. 1299-1305.

[5] M. S. Kavalekalam, M. G. Christensen, F. Gran, and J. B. Boldt, "Kalman filter for speech enhancement in cocktail party scenarios using a codebook-based approach," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016, pp. 191-195.

[6] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustic Speech Signal Processing*, vol. ASSP-27, no. 2, Apr. 1979, pp. 113-120.

[7] H. Gustafsson, S. Nordholm, and I. Claesson, "Spectral subtraction with adaptive averaging of the gain function," *6th European Conference on Speech Communication and Technology(Eurospeech'99)*, Budapest, Hungary, Sept. 1999, pp. 2599-2602.

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Trans. on Speech and Audio Processing*, vol. ASSP-32, no. 6, Dec. 1984, pp. 1109-1121.

[9] J. Lim and A. V. Oppenheim, "All‐pole modeling of degraded speech," *IEEE Trans. ASSP*, vol. 26, no. 3, 1978, pp. 197-210.

[10] X. Dang and T. Nakai, "Noise Reduction using Modified Phase Spectra and Wiener Filter," *2011 IEEE International Workshop on Machine Learning for Signal Processing*, Sept. 2011, pp. 1-5.

[11] F. Asano, S. Ikeda, M. Ogawa, H. Asoh, and N. Kitawaki, "Combined approach of array processing and independent component analysis for blind separation of acoustic signals," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 3, May 2003, pp. 204-215.

[12] J. Choi, "An Adaptive Speech Enhancement System Based on Noise Level Estimation and Lateral Inhibition," *ACTA Acustica United with Acustica*, vol. 93, no. 4, 2007, pp. 632-644.

[13] H. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition system under noisy conditions," *Proc. ISCA ITRW Workshop on Automatic Speech Recognition*, Paris, France, 2000.

저자 소개

최재승(Jae-Seung Choi)

1989년 조선대학교 전자공학과 공학사

1995년 일본 오사카시립대학 전자정보공학부 공학석사

1999년 일본 오사카시립대학 전자정보공학부 공학박사

2000년~2001년 일본 마쯔시타 전기산업주식회사(현, 파나소닉 주식회사) AVC사 연구원

2002년~2007년 경북대학교 디지털기술연구소 책임연구원

2007년~현재 신라대학교 스마트전기전자공학부 교수

※ 관심분야 : 음성신호처리, 신경회로망, 잡음제거, 음원분리 등