Journal of KIIT. Vol. 21, No. 3, pp. 93-98, Mar. 31, 2023. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2023.21.3.93

통계적 추정기 및 LPC 잔차잡음 억제에 기반한 음성강조 알고리즘의 SNR 효과

최 재 승*

The SNR Effect of Speech Enhancement Algorithm based on Statistical Estimator and Suppression of LPC Residual Noise

Jae-Seung Choi*

요 약

음성에 잡음이 중첩된 경우에 음성에 대한 선형예측 정밀도가 감소하게 되어 재생된 음성의 품질이 떨어진다. 따라서 본 논문에서는 다양한 잡음이 포함된 음성신호에 대하여 통계적 추정기에 의해서 개선된 선형예측 계수의 잔차신호를 사용하여 잡음제거를 목적으로 한 음성강조 알고리즘을 제안한다. 본 실험에서는 신호처리에 의한 SN비의 개선을 목적으로 하여 5종류의 배경잡음에 의하여 중첩된 음성신호를 사용하여 입력 SNR이 약 0dB~4dB인 경우에 대하여 실험을 진행하였으며, 각 잡음이 중첩된 음성에 대하여 출력 SNR이 평균 4.99(dB), 5.10(dB), 5.14(dB), 6.40(dB), 6.59(dB) 개선된 것을 알 수 있었다. 따라서 본 논문에서 제안한 알고리즘은 잡음억제의 SNR의 양호한 결과로부터 음성강조의 효과가 크다는 것을 명확히 할 수 있었다.

Abstract

When noise signal is included in speech, linear prediction accuracy for the speech is reduced, and thus the quality of the reproduced speech is degraded. Therefore, this paper proposes a speech enhancement algorithm for reducing the noise using the residual signal of the linear predictive coefficient improved by statistical estimator, for various noisy speech signals. In this experiment, for the purpose of improving the SNR by signal processing, this experiment was conducted for the case where the input SNRs are about 0dB to 4dB using noisy speech signal contaminated by 5 types of background noises. From these experiment results, it was confirmed that output SNRs improved by averages of 4.99(dB), 5.10(dB), 5.14(dB), 6.40(dB), and 6.59(dB) for each noisy speech signal. Therefore, it was clarified that the proposed algorithm has a good effect of speech enhancement from the SNR noise reduction results.

Keywords

linear predictive coding, residual signal, statistical estimator, speech enhancement, noise suppression

* 신라대학교 전기전자공학과 교수
- ORCID: https://orcid.org/0000-0002-5699-9701

· Received: Dec. 19, 2022, Revised: Jan. 10, 2023, Accepted: Jan. 13, 2023
· Corresponding Author: Jae-Seung Choi
Dept. of Electrical and Electronic Engineering, Silla University, 140 Baegyang-daero(Blvd), 700beon-gil(Rd), Sasang-gu, Busan, 46958 Korea
Tel.: +82-51-999-5608, Email: choijaes7@silla.ac.kr

Ⅰ. 서 론

음성은 사람의 의도를 전달하려는 한 가지 수단으로서 중요한 역할을 수행하고 있다. 기존의 음성신호 처리는 아날로그 파형 그 자체를 충실히 재현하려는 방법에 의하여 전달되었으나 디지털신호처리에 의한 신호의 고속처리 장치와 각종 정보단말기의 디지털화로 인하여, 음성신호를 전달하기 위한 디지털 음성신호의 부호화 방법 및 정보처리의 비중이 높아지고 있다[1]. 이러한 정보처리를 공학적으로 실현하는 경우에는 음성파형 그 자체를 처리하는 대신에 음성의 특징추출과 같이 음성파형에 특정한 전처리를 실행하여 고차의 정보처리를 실행하는 것이 일반적이다.

선형예측에 기초한 분석합성방식에 의하여 음원과 성도정보를 독립적으로 제어하는 방법이 종래부터 연구되고 있었지만, 이러한 방법들은 잡음이 포함되지 않은 음성을 처리대상으로서 검토되고 있었다. 잡음이 포함된 음성신호에 대해서 그대로 변환처리를 실행하면, 선형예측분석의 정밀도가 떨어져서 처리된 음성의 품질 열화가 발생하는 것을 예상할 수 있다[2]. 음성신호에 중첩된 잡음의 영향을 감소시키는 연구에 관해서는 음성부호화 및 음성인식의 분야에서 검토되고 있지만 이러한 방법들은 각각의 처리 내용에 특화된 것과 소요 연산량이 많다는 단점을 가지고 있다.

일상적인 생활환경에서 다양한 잡음이 발생하고 있으며 특히 휴대전화와 보청기 등을 적용한 마이크로폰 어레이에서는 희망하는 목적 신호인 음성 이외에 잡음이 혼입되면 일상 회화를 방해한다. 그러므로 쾌적한 음성통화 품질의 보전 및 잡음억압 기술을 실현할 목적으로 마이크로폰 어레이[3], 스펙트럼 차감(SS, Spectral Subtration)[4], 적응신호처리[5], 최소 이승오차 단시간 진폭스펙트럼 추정(MMSE-STSA, Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator)[6][7] 등의 잡음억제기술이 오래전부터 연구되어 왔다. 한편으로 높은 잡음 환경은 정상적인 음성에 의한 통신을 방해할 수 있는 요인이 된다. 이러한 문제요인을 해결하기 위해서는 잡음 레벨을 경감하는 것[8][9]과 잡음원의 제어가 우선적으로 해결되어야 한다. 또한 음성신호에 대한 신호잡음의 비율(SNR, Signal-to-Noise Ratio)의 저하는 전기통신에서 생기는 페이징 현상, 전송시스템의 불충분 등에 의한 통화품질의 열화와 관련 있는 공통된 문제이지만 음성파형신호의 SNR 및 통화품질을 간단하게 개선하는 것이 쉽지 않다. 신호처리에 의한 SNR의 개선은 잡음 환경 하의 음성통신뿐만 아니라 일반 음성통신의 통화품질에도 많은 도움이 된다. 음성의 경우에도 음성과 잡음의 성질의 차이에 착목하여 잡음레벨과 잔류잡음의 영향을 감소시켜 SNR을 개선하는 것이 가능하다[7][10].

음성에 잡음이 중첩되어 있으면 음성에 대한 선형예측의 정밀도가 감소하여 재생된 음성의 품질이 떨어진다. 따라서 본 논문에서는 이러한 관점에서 다양한 배경잡음을 사용하여 잡음에 의한 선형예측계수에 중점을 두고 연구를 진행한다. 다음으로 잡음억압 및 음성강조를 목적으로 하여 선형예측의 잔차신호 및 MMSE-STSA를 사용한 음성강조 알고리즘에 대해서 기술한다.

Ⅱ. 잡음제거를 위한 음성 강조법

이동전화, 디지털 보청기 및 음성인식 장치를 잡음이 많은 환경에서 사용할 경우에 주변잡음이 음성에 중첩되기 때문에 음성의 통화품질 열화 및 음성인식율의 열화를 일으킨다. 특히 음성인식 장치에서는 음성에 혼입한 배경잡음을 억제하여 목적으로 하는 음성신호 성분만을 추출하는 SS[4], MMSE-STSA[6][7] 방법이 제안되고 있다. 이러한 방법은 잡음스펙트럼 추정이 부정확하면 잡음억제 후에는 잡음이 잔류한다든가 잡음의 과대추정에 의해서 음성이 왜곡되어 음질이 열화하게 된다[4][6][7]. 근래에는 음성인식 및 음성강조의 분야에서 높은 잡음에서 성능향상을 위한 심층 신경회로망에 기초한 방법도 제안되고 있다[11]. 한편 적응디지털필터는 낮은 SNR 환경에 있어서 음성의 기본주기의 정확한 검출이 필요한 음성강조법이다[5].

본 논문에서는 MMSE-STSA 법에 의한 주파수영역의 음성강조법을 제안하며, 이 수법은 잡음이 중첩된 음성스펙트럼이 주어질 경우에 원래의 음성진

폭 스펙트럼과 추정된 음성진폭 스펙트럼과의 평균 이승오차를 최소로 하여 추정된 음성스펙트럼을 구한다[6][7]. MMSE-STSA 법은 진폭스펙트럼뿐만 아니라 위상에 관해서도 고려해야 하며, 음악잡음이 발생하기 어려운 수법으로 알려진 Ephraim 등에 의해서 제안된 MMSE-STSA가 있다[6].

원음성 $s(t)$에 주위 배경잡음 $\nu(t)$가 중첩하는 잡음중첩 음성을 $y(t)=s(t)+\nu(t)$와 같이 나타내며, 잡음중첩 음성의 주파수 스펙트럼을 $Y_f=S_f+\Omega_f$와 같이 정의한다. 여기에서, $f(0\le f\le N)$는 주파수번호를 나타내며 N은 프레임길이이다. 또한 Y_f, S_f, Ω_f는 각각 잡음중첩음성, 음성, 잡음의 주파수 스펙트럼을 나타내며 식 (1)과 같이 표현한다.

$$Y_f = Q_f e^{j\psi_f} \quad (1)$$
$$S_f = R_f e^{j\theta_f}$$
$$\Omega_f = W_f e^{j\xi_f}$$

원음성의 진폭스펙트럼 R_f와 추정음성진폭스펙트럼 $\hat{R}_f$와의 이승오차 평균값은 식 (2)와 같다.

$$E=\Xi[(R_f-\hat{R}_f)^2] \quad (2)$$

MMSE-STSA는 이승오차 평균치 E를 최소로 하는 추정 음성진폭스펙트럼을 구하는 방법이다. 추정음성 $\hat{S}(t)$는 잡음 중첩음성의 위상정보 ψ_f를 사용하여 $\hat{R}_f e^{j\psi_f}$를 역 푸리에변환함으로써 구한다.

III. 제안한 알고리즘에 의한 실험 및 고찰

본 논문에서 제안한 선형예측에 의한 부호화(LPC, Linear Predictive Coding) 수법은 음성파형으로부터 음성의 주파수 스펙트럼 구조를 명확히 하기 위하여, 어떤 시점에서 음성파형을 과거의 표본치의 가중치의 선형결합으로 나타내어 선형예측 오차의 이승평균을 최소화함으로써 선형예측계수를 추출 및 결정하는 방법이다[2].

각 프레임 i에 대하여 선형예측분석을 실행하여 식 (3)처럼 N차원의 LPC 계수 $a_i(k)$을 계산한다.

$$y(f)=\sum_{k=1}^{N} a_i(k)\, y(f-k) \quad (3)$$

산출한 LPC 계수 $a_i(k)$을 사용하여 식 (4)로부터 음성의 잔차신호 $\Gamma(f)$를 구한다.

$$\Gamma(f) = y(f) - \sum_{k=1}^{N} a_i(k)\, y(f-k) \quad (4)$$

본 논문에서는 선형예측계수에 개인성이 포함된다는 것에 착목하여 음성강조를 실행하기 위한 수법으로 그림 1에 나타내는 선형예측잔차를 사용한다. 일반적으로 음성의 선형예측모델에 있어서 현재의 음성신호는 선형예측계수와 사전의 음성신호로부터 예측되며, 이때 예측치와 실제 측정치와의 차이가 선형예측잔차로 불린다[12]. 그림 2는 잡음억압 및 음성강조 수법에 초점을 맞춘 LPC의 잔차신호 및 MMSE-STSA에 의하여 잡음을 제거한 음성강조 알고리즘을 나타낸다. 먼저 음성의 각 프레임에 대하여 선형예측분석에 의하여 구해진 LPC 계수를 사용하여 음성신호의 잔차신호가 구해진다. 구해진 LPC 잔차신호를 사용하여 본 논문에서 제안한 MMSE-STSA 법에 의하여 각각의 음성 구간에서 잡음의 스펙트럼 추정법을 사용하여 잡음을 제거한 후에 강조된 음성스펙트럼을 구하게 된다.

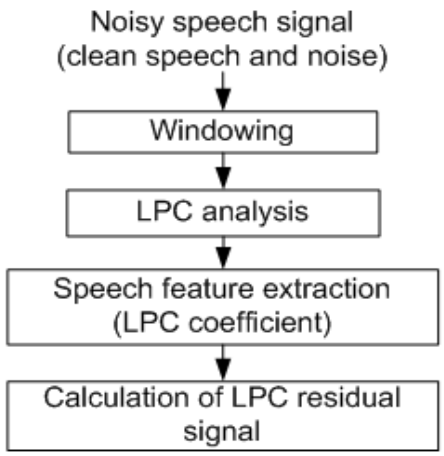

그림 1. LPC 잔차신호의 추출 과정
Fig. 1. Extraction procedure of LPC residual signal

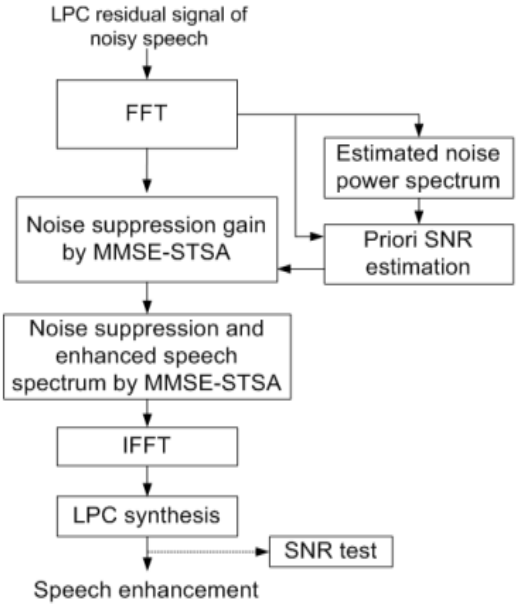

그림 2. MMSE-STSA 잡음제거의 블록도
Fig. 2. Block diagram of MMSE-STSA noise reduction

본 실험에서는 제안한 알고리즘의 성능을 분석하고 실험의 결과를 확인하기 위하여 NOIZEUS 데이터베이스를 사용한다[13]. NOIZEUS는 3명의 남성화자와 3명의 여성 화자(화자당 5개의 문장)에 의해서 생성한 30개의 IEEE 문장이 포함된 음성코퍼스이며, 원래 25 kHz에서 8 kHz로 다운 샘플링되었다.

실험에서 사용하는 잡음신호는 AURORA 데이터베이스에 포함된 자동차잡음(Car noise), 식당잡음(Restaurant noise), 공항잡음(Airport noise), 기차잡음(Train noise), 거리잡음(Street noise) 등을 사용했으며[13], 이러한 잡음을 NOIZEUS 데이터베이스의 음성신호에 중첩하여 제안한 알고리즘의 성능을 SNR을 사용하여 테스트하였다.

실험에 사용된 NOIZEUS 음성신호의 문장으로는 3명의 남성화자 “He knew the skill of the great young actress”(sp02.wav: MS1), “Her purse was full of useless trash” (sp03.wav: MS2), “Read verse out loud for pleasure”(sp04.wav: MS3), 3명의 여성화자로는 “The drip of the rain made a pleasant sound”(sp12.wav: FS1), “Smoke poured out of every crack”(sp13.wav: FS2), “Hats are worn to tea and not to dinner”(sp14.wav: FS3)이다[13].

실험에서 사용된 잡음이 중첩된 음성으로는 남성화자 3명 및 여성화자 3명으로 구성된 총 6개 문장으로 구성된 깨끗한 음성신호에 대해서 평균 입력의 SNR이 -0.17dB(Car noise), -0.24dB(Restaurant noise), -1.55dB(Airport noise), -3.37dB(Train noise), -3.75dB(Street noise)가 되도록 잡음신호를 중첩시켜 잡음 음성신호를 생성하였다.

본 논문에서 제안한 방식의 실험에서는 LPC 잔차신호에 대하여 MMSE-STSA를 사용해서 음성강조의 효과를 개선했으며, 이때 프레임 길이는 256, LPC 차수는 12이다. 본 실험에서는 제안한 알고리즘의 성능평가는 음성신호에 잡음신호가 얼마나 많이 부가되어 있는가를 나타내는 객관적인 음질평가 테스트 기법으로 알려진 출력 SNR을 이용하여 성능을 실험하였다.

표 1~5는 입력 잡음음성의 SNR(“Input SNR(dB)”)이 약 0dB～4dB인 경우에 본 논문에서 제안한 알고리즘에 의한 강조된 결과값(“Enhanced SNR(dB)”)과 개선된 결과값(“Improved SNR(dB)”)을 각각 나타내고 있다.

표 1과 표 2의 경우에 있어서, 표 1의 car noise가 중첩된 평균 SNR=-0.17dB인 음성에서는 “Enhanced SNR” 및 “Improved SNR”이 각각 4.82(dB), 4.99(dB) 개선되었으며, 표 2의 restaurant noise가 중첩된 평균 SNR= -0.24dB인 음성의 경우에는 각각 4.86(dB), 5.10(dB) 개선되었다. 표 3의 airport noise가 중첩된 평균 SNR= -1.55dB인 음성의 경우에는 “Enhanced SNR” 및 “Improved SNR”이 각각 3.59(dB), 5.14(dB) 개선되었으며, 표 4의 train noise가 중첩된 평균 SNR= -3.37dB인 음성의 경우에는 각각 3.03(dB), 6.40(dB) 개선되었다. 표 5의 street noise가 중첩된 평균 SNR= -3.75dB인 음성의 경우에는 “Enhanced SNR” 및 “Improved SNR”이 각각 2.84(dB), 6.59(dB) 개선되었다. 실험결과로부터 입력되는 잡음의 양이 많아지면 질수록 제안한 알고리즘에 의하여 출력 SNR에 대한 개선된 효과가 증대되는 것을 알 수 있었다.

특히 표 4의 train noise(평균 입력 SNR=-3.37dB)와 표 5의 street noise(평균 입력SNR=-3.75dB)가 중첩된 음성에서 각각 “Improved SNR”이 6.40(dB), 6.59(dB) 개선되어 본 알고리즘의 효과가 크다는 것을 명확히 할 수 있었다.

표 1. Car noise가 중첩된 음성의 실험 결과
Table 1. SNR results using car noise

Speech	Input SNR (dB)	Enhanced SNR (dB)	Improved SNR (dB)
MS1	-0.17	4.98	5.15
MS2	-0.20	5.97	6.17
MS3	-0.12	4.45	4.57
FS1	-0.18	4.16	4.34
FS2	-0.15	6.13	6.28
FS3	-0.17	3.25	3.42
Average	-0.17	4.82	4.99

표 2. Restaurant noise가 중첩된 음성의 실험 결과
Table 2. SNR results using restaurant noise

Speech	Input SNR (dB)	Enhanced SNR (dB)	Improved SNR (dB)
MS1	-0.28	4.69	4.97
MS2	-0.31	5.59	5.90
MS3	-0.23	4.57	4.80
FS1	-0.21	4.30	4.51
FS2	-0.21	6.16	6.37
FS3	-0.19	3.84	4.03
Average	-0.24	4.86	5.10

표 3. Airport noise가 중첩된 음성의 실험 결과
Table 3. SNR results using airport noise

Speech	Input SNR (dB)	Enhanced SNR (dB)	Improved SNR (dB)
MS1	-1.66	3.35	5.01
MS2	-1.80	4.02	5.82
MS3	-1.39	3.36	4.75
FS1	-1.62	3.61	5.23
FS2	-1.41	4.40	5.81
FS3	-1.40	2.81	4.21
Average	-1.55	3.59	5.14

표 4. Train noise가 중첩된 음성의 실험 결과
Table 4. SNR results using train noise

Speech	Input SNR (dB)	Enhanced SNR (dB)	Improved SNR (dB)
MS1	-3.51	3.31	6.82
MS2	-3.98	3.17	7.15
MS3	-3.26	2.72	5.98
FS1	-3.38	3.09	6.47
FS2	-3.01	3.52	6.53
FS3	-3.07	2.37	5.44
Average	-3.37	3.03	6.40

표 5. Street noise가 중첩된 음성의 실험 결과
Table 5. SNR results using street noise

Speech	Input SNR (dB)	Enhanced SNR (dB)	Improved SNR (dB)
MS1	-3.93	3.15	7.08
MS2	-4.41	3.14	7.55
MS3	-3.68	2.41	6.09
FS1	-3.74	2.56	6.3
FS2	-3.43	3.40	6.83
FS3	-3.32	2.35	5.67
Average	-3.75	2.84	6.59

Ⅳ. 결 론

본 논문에서는 잡음이 포함된 음성신호에 대해서 신호처리에 의한 SN비의 개선을 목적으로 하여, 선형예측분석에 의해서 구한 선형예측의 잔차신호에 MMSE-STSA법을 적용한 연구를 진행하였다. 본 실험에서는 5종류의 배경잡음에 의하여 중첩된 음성신호를 사용하여 입력 SNR이 약 0dB~4dB인 경우에 대하여 잡음을 억제하기 위한 실험을 진행하였다. 실험에서 각 잡음이 중첩된 음성신호에 대하여 출력 SNR이 평균적으로 car noise에서는 4.99(dB), restaurant noise에서는 5.10(dB), airport noise에서는 5.14(dB), train noise에서는 6.40(dB), street noise에서는 6.59(dB) 개선된 것을 알 수 있었다. 따라서 본 논문에서 제안한 알고리즘은 특히 입력잡음의 양이 많은 street noise에 대해서 출력 SNR의 개선효과가 상당히 증대됨으로써 잡음억제 및 음성강조의 효과가 크다는 것을 명확히 할 수 있었다.

References

[1] S. Cheng, H. Zhang, and G. Hua, "Speech Enhancement in Low SNR Environments by Designing a Time-Frequency Binary Mask", In 2018 IEEE 23rd International Conference on Digital Signal Processing, Shanghai, China, pp. 1-5, Nov. 2018. https://doi.org/10.1109/ICDSP.2018.8631645.

[2] J. S. Choi, "Gender Recognition for Speaker in Colored Noise by Speech/Non-speech

Discrimination", The Journal of KIIT, Vol. 10, No. 11, pp. 63-68, Nov. 2012.

[3] Y. Kaneda and J. Ohga, "Adaptive microphone-array system for noise reduction", IEEE Trans. Acoust. Speech Signal Process., Vol. 34, No. 6, pp. 1391-1400, Dec. 1986. https://doi.org/10.1109/TASSP.1986.1164975.

[4] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", IEEE Trans. on Acoustic Speech Signal Processing, Vol. 27, No. 2, pp. 113-120, Apr. 1979. https://doi.org/10.1109/TASSP.1979.1163209.

[5] M. R. Sambur, "Adaptive noise canceling for speech signals", IEEE Trans. Acoust. Speech Signal Process., Vol. 26, No. 5, pp. 419-423, Oct. 1978. https://doi.org/10.1109/TASSP.1978.1163137.

[6] 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. 32, No. 6, pp. 1109-1121, Dec. 1984. https://doi.org/10.1109/TASSP.1984.1164453.

[7] S. Y. Jung and K. S. Bae, "Feature Extraction through the post processing of WFBA based on MMSE-STSA for Robust Speech Recognition", Proc. of the Acoustical Society of Korea Conference, Vol. 23, No. 2, pp. 39-42, 2004.

[8] Y. G. Sun, Y. M. Hwang, I. Sim, and J. Y. Kim, "De-noising in Power Line Communication Using Noise Modeling Based on Deep Learning", The Journal of IIBC, Vol. 18, No. 4, pp. 55-60, Aug. 2018. https://doi.org/10.7236/JIIBC.2018.18.4.55.

[9] I. H. Jee, "A Study on Noise Removal Using Over-sampled Discrete Wavelet Transforms", The Journal of IIBC, Vol. 19, No. 1, pp. 69-75, Feb. 2019. https://doi.org/10.7236/JIIBC.2019.19.1.69.

[10] M. Souden, J. Benesty, and S. Affes, "On the Global Output SNR of the Parameterized Frequency-Domain Multichannel Noise Reduction Wiener Filter", IEEE Signal Processing Letters, Vol. 17, No. 5, pp. 425-428, May 2010. https://doi.org/10.1109/LSP.2010.2042520.

[11] D. S. Park, J. I. Bang, H. J. Kim, and Y. J. Ko, "A Study on the Gender and Age Classification of Speech Data Using CNN", The Journal of KIIT, Vol. 16, No. 11, pp. 11-21, Nov. 2018. https://doi.org/10.14801/jkiit.2018.16.11.11.

[12] K. V. Sorensen and S. V. Andersen, "Speech enhancement with natural sounding residual noise based on connected time-frequency speech presence regions", EURASIP Journal on Applied Signal Process, Vol. 2005, No. 1, pp. 2954-2964, Jan. 2005. https://doi.org/10.1155/ASP.2005.2954.

[13] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms", Speech Communication, Vol. 49, No. 7, pp. 588-601, Jul. 2007. https://doi.org/10.1016/j.specom.2006.12.006.

저자소개

최 재 승 (Jae-Seung Choi)

1989년 : 조선대학교 전자공학과 (공학사)
1995년 : 일본 오사카시립대학 전자정보공학부(공학석사)
1999년 : 일본 오사카시립대학 전자정보공학부(공학박사)
2000년 ~ 2001년 : 일본 마쯔시타 전기산업주식회사(현, 파나소닉) AVC사 연구원
2002년 ~ 2007년 : 경북대학교 디지털기술연구소 책임연구원
2007년 ~ 현재 : 신라대학교 전기전자공학과 교수
관심분야 : 음성인식, 음성강조, 잡음제거, 음원분리