ISSN 1975-8359(Print) / ISSN 2287-4364(Online)
The Transactions of the Korean Institute of Electrical Engineers Vol. 64, No. 12, pp. 1756~1760, 2015
http://dx.doi.org/10.5370/KIEE.2015.64.12.1756

음성 강화를 위한 a priori SNR 추정기반 적응 바람소리 저감 방법

An Adaptive Wind Noise Reduction Method Based on a priori SNR Estimation for Speech Eenhancement

서 지 훈* · 이 석 필†
(Ji-Hun Seo · Seok-Pil Lee)

Abstract - This paper focuses on a priori signal to noise ratio (SNR) estimation method for the speech enhancement. There are many researches for speech enhancement with several ambient noise cancellation methods. The method based on spectral subtraction (SS) which is widely used in noise reduction has a trade-off between the performance and the distortion of the signals. So the need of adaptive method like an estimated a priori SNR being able to making a high performance and low distortion is increasing. The decision directed (DD) approach is used to determine a priori SNR in noisy speech signals. A priori SNR is estimated by using only the magnitude components and consequently follows a posteriori SNR with one frame delay. We propose a modified a priori SNR estimator and the weighted rational transfer function for speech enhancement with wind noises. The experimental result shows the performance of our proposed estimator is better Perceptual Evaluation of Speech Quality scores (PESQ, ITU-T P.862) compare to the conventional DD approach-based systems and different noise reduction methods.

Key Words : Noise reduction, A priori SNR, Speech enhancement

1. 서 론

최근 음성인식 기술의 발단과 함께 스마트 폰 등의 일부 제품에서 음성인식 기능을 사용하면서 점차 음성인식에 대한 관심이 높아지고 있다. 실제 환경에서의 음성 신호는 대부분 다양한 종류의 배경 잡음을 포함하며 이러한 배경 잡음은 음성 인식 시스템의 성능을 저하시키는 주요한 원인이 된다. 또한 실제 환경에서 통화의 품질을 높이기 위해 바람소리의 제거가 최근 이슈가 되고 있다. 이러한 잡음 제거의 중요성으로 다양한 잡음 제거 알고리즘이 연구되어 왔다[1-4]. 잡음을 제거하고 음성신호를 향상시키는 기술은 음질을 높일 뿐 아니라 음성 인식률을 개선하고 음성의 명료도를 높일 수 있다.

잡음을 제거하는 기술은 크게 두 가지로 나눌 수 있다. 잡음을 추정하여 원신호에서 빼주는 스펙트럼 차감법(spectral subtraction, SS)[6]기반의 방법과 음성신호를 추정하여 원신호를 갱신하는 wiener filter[7]기반의 방법이 있다. 스펙트럼 차감법은 배경잡음을 제거하는 가장 간단하고 효과적인 방법이지만, 차감하는 과정에서 발생하는 잔류잡음은 음성인식의 성능을 크게 저하시키며 사람의 귀에 거슬리는 소리를 발생시킨다. 또한 낮은 SNR 환경에서 음성왜곡이 심하다는 문제점을 가지고 있다. 이러한 문제점을 개선하기 위해 파라미터를 이용하는 방법이 연구되었다[3. 5]. 파라미터를 이용하는 방법은 잡음의 크기를 이용하여 SNR을 추정하거나[3], 마스킹 효과를 이용하여 마스킹 임계치를 계산한 후 이 임계치값을 가지고 파라미터를 추출하는 방법이 있다[5]. 하지만 파라미터를 이용한 방법은 파라미터 결정에 따라 결과가 크게 변하는 문제가 있다. 결정 파라미터 값을 크게 하면 잡음제거 성능은 좋아져 잔류잡음은 감소하지만 음성이 왜곡 될수 있다. 반대로 파라미터 값을 작게 하면 음성왜곡은 감소하지만 배경잡음 제거 성능은 저하 된다.

따라서 파라미터를 이용한 방법에서 파라미터를 결정하는 것은 매우 중요한 문제이다. wiener filter는 잡음에 원하는 음성신호를 추정하여 잡음을 제거하는 필터 이다. 이와 유사한 방법에는 Matched filter[8], Kalman filter[2], least mean square (LMS)[9], recursive least square (RLS)[10]등이 있다. 하지만 이러한 적응 필터는 잡음이 섞인 신호에서 음성신호를 추정하는 것이 어렵고 불안정한 환경에는 적용시키기 어렵다는 단점이 있다.

본 논문에서는 다양한 잡음들 중에서 특히 바람소리를 제거하여 통화의 품질 및 인식성능을 향상시키기 위해 스펙트럼 차감법 기반의 파라미터를 이용한 방법들의 단점을 보완하여 음성왜곡을 최소화하고 낮은 SNR환경에서도 좋은 성능을 보이는 방법을 제안하고자 한다. 제안하는 방법은a priori SNR을 추정하여 원신호에서 잡음을 제거하는 방법으로, 음성신호 추정값을 선형예측하고, 주파수별로 a priori SNR의 가중치를 다르게 하여 주변 환경의 변화에 쉽게 적응하여 낮은 SNR에서도 좋은 성능을 기대할

† Corresponding Author : Dept. of Media software, Sangmyung University, Korea
E-mail : esprit@smu.ac.kr
* Dept. of Computer science, Sangmyung University, Korea
Received : November 2, 2015; Accepted : November 27, 2015

수 있다. 제안한 방법의 성능 평가를 위해 객관적인 음질평가 테스트 ITU-T P.862 perceptual evaluation of speech quality (PESQ)[11]를 실시하여 다양한 잡음환경에서 기존의 방법들과 비교하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 방법과 제안하는 방법에 대해 기술하고, 3장에서는 기존 연구와 실험 결과를 비교한다. 마지막으로 4장에서는 제안한 방법의 결론에 대해 논의한다.

2. 기존 연구 및 제안하는 방법

본 장에서는 본 논문에서 제안하는 방법의 기초가 되는 스펙트럼 차감법에 대해 설명 하고, 기존의 방법의 문제점을 보완한 방법에 대해 기술 한다. 단일 마이크로폰으로 얻은 음성 신호와 잡음 신호를 각각 x(t), n(t)라고 할 때, 음성신호와 잡음 신호가 혼합된 혼합 신호 y(t)는 두 신호의 합으로 다음과 같이 나타낼 수 있다.

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

이 신호를 시간축으로 윈도우(Hamming window)를 씌어서 단구간 푸리에 변환(FFT)을 사용하였을 때, 입력신호와 잡음신호가 서로 상관이 없다고 가정한다면, 혼합신호의 파워스펙트럼 Y(k)는 다음과 같이 나타 낼 수 있다.

$$|Y(k)|^2 = |X(k)|^2 + |N(k)|^2 \tag{2}$$

X(k)는 음성신호의 파워스펙트럼이며, N(k)는 잡음신호의 파워스펙트럼 이다.

스펙트럼 차감법은 잡음 추정값을 계산하는 과정과 추정 잡음을 이용하여 잡음을 제거하는 과정이 있다. 잡음 추정값을 계산하는 과정은 목적 신호의 단구간 진폭 스펙트럼의 합을 윈도우 크기 N으로 나눈 평균을 계산 하여 잡음 추정값 $|\hat{N}(k)|$을 구하며, 식 (3)과 같이 계산하여 구할 수 있다. 잡음을 제거하는 과정은 목적 신호에서 잡음 추정값을 차감하여 음성 신호 추정값 $|\hat{X}(k)|$을 구하며, 식 (4)와 같다.

$$|\hat{N}(k)| = \frac{1}{N}\sum_{i=1}^{N}|N_i(k)| \tag{3}$$

$$|\hat{X}(k)| = |Y(k)| - |\hat{N}(k)| \tag{4}$$

식 (4)로부터 계산된 음성 신호 추정값 $|\hat{X}(k)|$에 입력 신호의 위상을 부가하여 역 푸리에변환을 수행함으로써 향상된 음성신호를 얻을 수 있다. 기존의 스펙트럼 차감법에 대한 구조도를 그림 1에 나타낸다.

기존의 방법은 낮은 SNR 환경에서 높은 성능을 기대하기 힘들며, 잡음 추정값이 고정됨으로써 변화하는 잡음에 적응하기 어

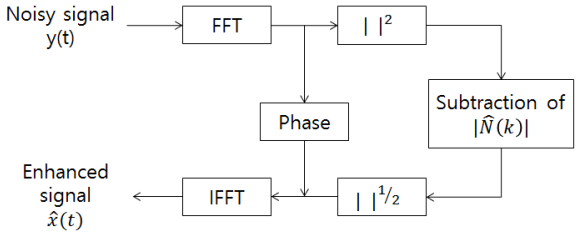


그림 1 기존의 스펙트럼 차감법
Fig 1 Architecture of spectrum subtraction

렵다는 문제를 가지고 있다. 이러한 문제를 해결하기 위해 본 논문에서는 낮은 SNR에서도 환경에 쉽게 적응할 수 있도록 a priori SNR을 추정하여 소음을 저감시키는 방법을 제안 한다.

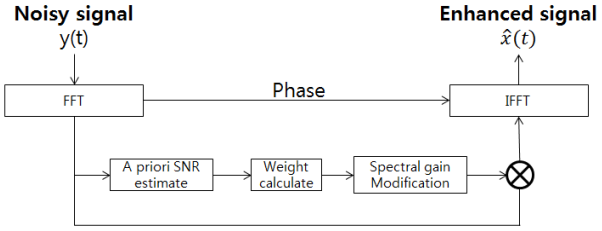


그림 2 제안하는 방법
Fig. 2 Proposed wind reduction method

그림 2는 본 논문에서 제안하는 방법의 구조도 이다. 제안하는 방법은 a priori SNR을 추정한 후 추정한 값을 이용하여 가중치 값을 계산하고, 추정된 a priori SNR과 가중치 값을 noisy 신호에 곱하여 잡음을 제거하는 방법이다. 파라미터를 이용하는 잡음 제거 방법에서는 a priori SNR과 a posteriori SNR을 이용한다. a priori SNR은 음성 신호와 잡음 신호의 파워 스펙트럼의 비이고, a posteriori SNR은 혼합신호와 잡음신호의 비로 다음과 같이 정의 된다.

$$SNR_{prio}(p,k) = \frac{E\{|X(p,k)|^2\}}{E\{|N(p,k)|^2\}} \tag{5}$$

$$SNR_{post}(p,k) = \frac{|Y(p,k)|^2}{E\{|N(p,k)|^2\}} \tag{6}$$

X(p, k), N(p, k) 그리고 Y(p, k)는 음성신호, 잡음신호, 혼합신호의 p번째 프레임의 k번째 주파수 성분을 나타내고, E{}는 평균값 계산을 나타 낸다. instantaneous SNR은 다음과 같이 정의 된다.

$$SNR_{inst}(p,k) = \frac{|Y(p,k)|^2 - E\{|N(p,k)|^2\}}{E\{|N(p,k)|^2\}} \tag{7}$$

instantaneous SNR은 SS기법에 기반하여 짧은 순간의 a priori SNR을 추정할 수 있으며, a priori SNR 추정 방법을 평

가하는데 유용하다.

하지만, 우리는 혼합 신호 X(p, k)만 관측할 수 있으며, 음성 신호와 잡음 신호는 알지 못하므로 a priori SNR과 a posteriori SNR을 추정해야 한다. SS방법에 따르면 혼합신호의 앞쪽 단구간을 잡음이라고 가정할 수 있으므로 우리는 잡음스펙트럼은 추정할 수 있다. 이 추정값을 통해 우리는 a posteriori SNR은 쉽게 추정 가능하다. 식 (2), (5), (6)을 이용하여 정리해 보면, 다음과 같이 나타낼 수 있다.

$$SNR_{prio}(p,k) = \frac{E\{|Y(p,k)|^2 - |N(p,k)|^2\}}{E\{|N(p,k)|^2\}} = SNR_{post} - 1 \tag{8}$$

하지만 a priori SNR을 식 (6)과 같이 추정을 하게 되면 a priori SNR이 a posteriori SNR의 형태를 따라가게 되는 문제점이 발생하게 되며, 고정된 값 1을 차감하게 되어 기존의 SS의 문제점도 해결하지 못한다. 이러한 문제점을 해결하기 위한 a priori SNR을 추정하는 방법 중 가장 널리 알려진 방법이 decision directed(DD) 접근법이다[4]. DD 접근법은 다음과 같이 나타낸다.

$$SNR_{prio}(p,k) = E\{\alpha * \frac{|\hat{X}(p,k)|^2}{E\{|\hat{N}(p,k)|^2\}} + (1-\alpha) * \mathrm{SNR}_{inst}(p,K)\} \tag{9}$$

하지만, 위 식의 표현은 실제로 계산할 수 없기 때문에 다음 식 (8)과 같이 재귀적으로 근사하여 a priori SNR을 계산한다.

$$SNR_{prio}(p,k) = \alpha * \frac{|\hat{S}(p-1,k)|^2}{E\{|N(p-1,k)|^2\}} + (1-\alpha) * \max(SNR_{post}(p,k)-1,0) \tag{10}$$

α는 가중치 파라미터로 0에서 1사이의 값을 갖는다. $|\hat{S}(p-1,k)|^2$은 이전 프레임에서 추정된 음성 신호의 스펙트럼 값이다. 기존 DD 접근법을 이용한 방법은 이전 프레임에서 추정된 음성 신호의 스펙트럼을 이용하기 때문에 한 프레임의 시간지연을 가지며, a posteriori SNR을 따르는 경향이 있다. 이러한 문제를 해결하기 위해 제안하는 방법은 a posteriori SNR 추정값 합리적 전달함수를 이용하여 a priori SNR을 추정하고, 추정한 a priori SNR을 이용하여 가중치 값을 계산하여 잡음제거 정도를 조절하도록 한다. 본 논문에서 제안하는 a priori SNR은 다음과 같다.

$$SNR_{prio}(p,k) = \alpha * \frac{|\hat{S}(p-1,k)|^2}{E\{|N(p-1,k)|^2\}} + \sum_{n=1}^{p-1}(1-\alpha)^n * \max(SNR_{post}(p-n,k)-1,0) \tag{11}$$

식 (11)과 같이 계산하면 시간축 상에서 가까운 프레임의 추정값에 더 많은 영향을 받으므로 주변 환경에 적응하면서 a priori SNR값을 추정할 수 있게 되며, 이전 a priori SNR의 형태를 따르는 문제를 어느 정도 해결할 수 있다.

그 후, 추정한 a priori SNR을 이용하여 잡음 제거 필터를 계산 하고, 잡음 제거 필터와 혼합 신호를 곱해 잡음을 제거 한다. 잡음 제거필터 H(p,k)는 다음과 같다.

$$H(p,k) = \frac{SNR_{prio} * \mu}{1 + SNR_{prio} * \mu} \tag{12}$$

μ는 가중치 파라미터 이고, μ의 값은 추정된 a priori SNR값을 이용하여 다음과 같이 계산 한다.

$$\mu = \frac{\sqrt{(SNR_{prio})^2 + |SNR_{prio}|}}{|SNR_{prio}|} \tag{13}$$

A priori SNR의 값이 크게 추정이 되면 음성신호 크기가 크고 잡음신호는 작기 때문에 가중치 값도 커져야 하며, a priori SNR 값이 작으면 음성신호 크기가 작기 때문에 가중치 값도 작아져야 한다.

잡음제거 필터를 곱하여 구한 향상된 음성 신호는 다음과 같이 계산 된다.

$$\hat{S}(p,k) = Y(p,k) * H(p,k) \tag{14}$$

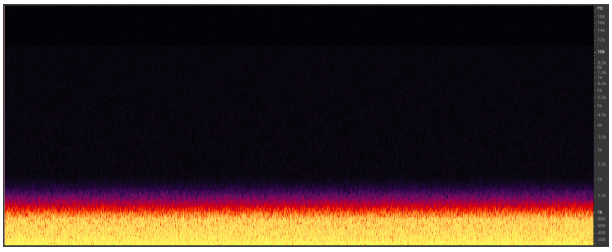
식 (14)를 이용하여, 역 이산 푸리에 변환(IDFT)을 하여, 음성이 강화 된 신호를 복원 할 수 있다.

3. 실험 결과

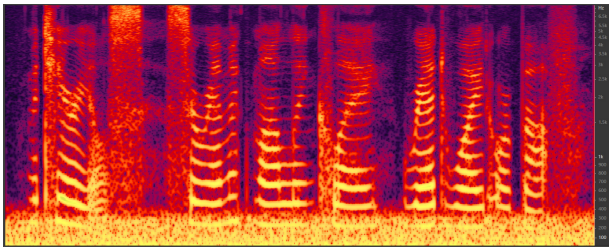
본 논문에서 제안한 방법의 성능을 평가하기 위해 객관적인 음질 평가방법인 PESQ를 사용 하였다. 음성 데이터는 TIMIT 데이터베이스[12]에서 SA타입으로 발음하는 남녀 20명의 데이터와 core test set 192개를 사용 하여 총 232개를 사용 하였다. 잡음 데이터는 녹음한 바람소리 데이터 10개를 사용 하였다. 합성한 신호는 샘플링 레이트16kHz, 16bit, 윈도우 크기는 1024로 하여 50% 오버랩 하였고, SNR에 따른 성능을 확인하기 위해 SNR -10에서 10dB까지 5dB씩 증가시키며 진행하였다.

그림 3은 바람소리잡음, 합성신호 그리고 처리 후 신호에 대한 스펙트로그램이다. (a)는 비교 실험에 사용한 바람소리 잡음 중 하나 이다. 스펙트로그램을 보면 1000Hz 이하의 대역에서 큰 에너지를 가지고 있음을 확인할 수 있다. (c)를 보면 본 논문에서 제안한 방법으로 바람소리 저감을 하였을 때, 저대역에 있던 잡음이 많이 저감되었음을 확인할 수 있으며 음성 신호가 더 뚜렷하게 나타나고 있다.

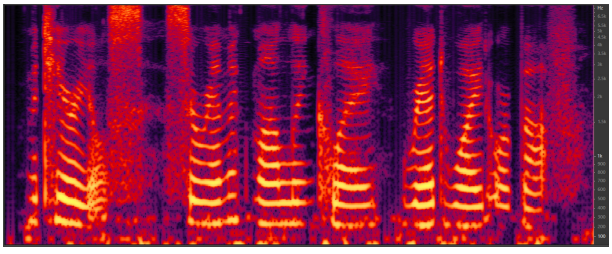
표 1은 기존의 SS 기반 연구, STSA 추정 기반 연구 그리고



(a) 바람 소리 잡음



(b) 합성된 noisy 신호



(c) 제안한 방법 처리 후 신호

그림 3 잡음 신호, 합성 신호, 처리 후 신호 스펙트로그램
Fig. 3 The spectrogram of signal

표 1 바람소리 잡음 향상량
Table 1 The PESQ improvement result of wind noise

SNR	-10dB	-5dB	0dB	5dB	10dB
SS[13]	+0.15	+0.19	+0.26	+0.32	0.36
STSA[14]	-	+0.44	+0.42	+0.41	+0.43
LSA[15]	-	+0.46	+0.46	+0.47	+0.5
WE[16]	-	+0.47	+0.48	+0.51	+0.55
Proposed	+0.81	+0.85	+0.89	+0.85	+0.70

제안하는 방법으로 바람소리를 저감시킨 후의 PESQ 변화량을 나타낸 표이다. 표 1의 실험의 목적은 SS 기반, STSA 추정 기반과 변형된 STSA 추정 기반 연구들과의 비교 이다. 비교한 기존 연구는 SS[13], short time spectral amplitude 추정 (STSA)[14], log spectral amplitude (LSA)[15] 추정, weighted euclidean (WE)[16] 기반 추정 방법이다. 기존 연구의 결과 수치는 논문에 나와 있는 수치를 그대로 사용하였고, 본 논문에서 제안한 방법의 수치는 실험에 사용한 232명의 PESQ의 평균값을 사용하였다. 결과를 보면 실험을 진행한 모든 SNR 환경에서 기존의 방법보다 더 좋은 음질 향상이 있음을 확인할 수 있다. 다른 유형의 바람소리에 대한 실험결과도 유사한 향상량을 보였다.

4. 결 론

실제 환경에서의 음성 신호는 대부분 다양한 종류의 배경 잡음을 포함하며 이러한 배경 잡음은 음성 인식 시스템의 성능을 저하시키는 주요한 원인이 된다. 또한 실제 환경에서 통화의 품질을 높이기 위해 바람소리의 제거가 최근 이슈가 되고 있다. 본 논문에서는 바람소리를 효과적으로 제거하기 위해 a priori SNR을 추정하고 추정한 a priori SNR값으로 가중치값을 계산하여 잡음을 제거하는 방법을 제안하였다. 제안한 방법의 성능을 평가하기 위해 기존의 스펙트럼 차감법과 PESQ을 이용하여 성능을 측정 비교하였고, 실험 결과 제안하는 방법은 기존의 스펙트럼 차감법, STSA기반 방법 보다 우수한 성능을 보였고, PESQ 향상량을 보았을때 큰 향상효과가 있음을 확인 할 수 있었다. 또한, 낮은 SNR 환경에서도 우수한 성능을 보이고 있어 바람소리 저감 효과가 있음을 확인 할 수 있다.

References

[1] K. Daqrouq, I. N. Abu-Isbeih, M. Alfauri,"Speech signal enhancement using neural network and wavelet transform", 2009 6th International Multi-Conference on Systems, Signals and Devices, pp. 1-6, 2009.

[2] Y. Shao, C.H. Chang, "A Kalman filter based on wavelet filter-bank and psychoacoustic modeling for speech enhancement", 2006 IEEE International Symposium on Circuits and Systems, ISCAS 2006. Proceedings, May, 2006.

[3] Steven F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", IEEE Transactions on Signal Processing, 27(2), pp. 113-120, 1979

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator", IEEE.Transactions in Acoust., Speech, Signal Process., vol. 32, no. 6, pp. 1109-1121, Dec. 1984.

[5] Sinha, Deepen, and Ahmed H. Tewfik. "Low bit rate transparent audio compression using adapted wavelets", Signal Processing, IEEE Transactions on 41.12 (1993): 3463-3479.

[6] M. Bhatnagar, "A modified spectral subtraction method combined with perceptual weighting for speech enhancement", Master's thesis, University of Texas at Dallas, pp. 1-10, 2003.

[7] O. Cappe, “Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor”, IEEE Trans. Speech Audio Processing, 2 (2), pp. 346-349, 1994.
[8] J. Freudenberger, S. Stenzel, "Blind Matched Filtering for Speech Recording in Uncorrelated Noise", International Workshop on Acoustic Signal Enhancement, pp. 1-4, September 2012.
[9] Haykin, Simon, and Bernard Widrow. “Least-mean-square adaptive filters.” Vol. 31. John Wiley & Sons, 2003.
[10] Guopin, H., Wei, Z., Qin, Z. “Improvement of audio noise reduction system based on RLS algorithm.” Computer Science and Network Technology (ICCSNT), 2013 3rd International Conference on. IEEE, 2013.
[11] ITU-T P.862, Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-band Telephone Networks and Speech Codecs, 2001.
[12] TIMIT: acoustic-phonetic continuous speech corpus. Linguistic Data Consortium, 1993.
[13] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction”, IEEE Trans. Acoust., Speech, Signal Processing, Vol. 27, No. 2, pp. 113-120, 1979.
[14] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean square error short-time spectral amplitude estimator,” IEEE Transactions in Acoust., Speech, Signal Process., vol. 32, no. 6, pp. 1109-1121, Dec. 1984.
[15] Y. Ephraim and D. Malah, “Speech Enhancement Using a Minimum Mean-Sqaure Error Log-Spectral Amplitude Estimator”, IEEE Transactions on Acoustic, Speech and Signal Processing, vol. 33, no. 2, 1985, pp. 443-445.
[16] P. C. Loizou, “Speech Enhancement based on Perceptually Motivated Bayesian Estimators of the Magnitude Spectrum”, IEEE Transactions on Speech and Audio Processing, vol. 13, no. 5, 2005, pp. 857-869.

저 자 소 개

서 지 훈 (Ji-Hun Seo)

2014년 상명대학교 디지털미디어학과 이학사. 2014년～현재 상명대학교 컴퓨터과학과 석사과정
〈주관심 분야〉 오디오 신호처리, 패턴인식

이 석 필 (Seok-Pil Lee)

1990년 연세대학교 전기공학과 공학사
1992년 연세대학교 전기공학과 공학석사
1997년 연세대학교 전기공학과 공학박사
1997년～2002년 대우전자 영상연구소 선임연구원. 2002년～2012년 KETI 디지털미디어연구센터 센터장. 2010년～2011년 미국 Georgia Tech. 방문 연구원. 2012년～현재 상명대학교 미디어소프트웨어학과 교수
〈주관심 분야〉 멀티미디어 검색, 방송통신시스템, 인공지능