지능형 TV의 음성인식을 위한 참조 잡음 기반 음성개선

정상배*

Reference Channel Input-Based Speech Enhancement for Noise-Robust Recognition in Intelligent TV Applications

Sangbae Jeong*

이 논문은 지식경제부와 한국산업기술평가관리원의 정보통신연구개발사업[10041610, 인식센서융합 기반 실환경 하에서 임의의 사용자 30명에 대한 인식률 99%에 근접하는 사용자의 신원과 행위 및 위치 정보 인식 기술 개발 (2차년도)]의 지원을 받아 수행되었음

요 약

본 논문에서는 지능형 TV의 음성인터페이스를 위한 잡음제거 시스템에 대해서 제안한다. 음성인식 성능 저하에 매우 나쁜 영향을 주는 TV 소리를 제거하기 위해서 TV 소리 자체를 참조 잡음으로 하는 잡음제거 알고리즘이 구현된다. 제안된 알고리즘에서 TV 스피커와 다채널 장비간의 전달함수를 추정한다. 그 후, 위너 필터를 동작시키기 위해서 잡음의 전력 스펙트럼이 추정된다. 추가적으로 후처리 과정이 적용되어 잔존 잡음을 제거한다. 실험의 의해서 제안된 알고리즘이 5 dB 입력 SNR에서 88 %의 음성인식률을 나타내었다.

ABSTRACT

In this paper, a noise reduction system is proposed for the speech interface in intelligent TV applications. To reduce TV speaker sound which are very serious noises degrading recognition performance, a noise reduction algorithm utilizing the direct TV sound as the reference noise input is implemented. In the proposed algorithm, transfer functions are estimated to compensate for the difference between the direct TV sound and that recorded with the microphone installed on the TV frame. Then, the noise power spectrum in the received signal is calculated to perform Wiener filter-based noise cancellation. Additionally, a postprocessing step is applied to reduce remaining noises. Experimental results show that the proposed algorithm shows 88% recognition rate for isolated Korean words at 5 dB input SNR.

키워드

잡음제거, 음성인식, 위너필터

Key word

noise reduction, speech recognition, Wiener filter

* 정회원 : 경상대학교 (jeongsb@gnu.ac.kr)

접수일자 : 2012. 07. 30
심사완료일자 : 2012. 09. 05

Open Access http://dx.doi.org/10.6109/jkiice.2013.17.2.280

Ⅰ. 서 론

최근 기계-인간 사이의 인터페이스에 대한 많은 연구가 수행되고 있다. 그 중 음성인터페이스는 가장 편리하고, 자연스러운 의사소통 방식으로 알려져 있다. 과거에 기술적인 한계에 부딪혀 제한적인 환경에서 이용 되었던 것에 비하여, 최근에는 지능형 로봇, 스마트폰, 자동차 네비게이션 등의 다양한 곳에서 활용되고 있다. 특히, 최근 부각되고 있는 지능형 TV에서의 음성인터페이스 활용은 TV에서 활용 가능한 방대한 어플리케이션 및 컨텐츠 검색 을 직관적이면서도 편리하게 수행할 수 있다. 향후에 TV에서의 음성인터페이스가 본격적으로 상용화된다면 기존의 TV 리모컨이 필요 없게 될 것이다.

그러나 인간과 TV 사이의 인터페이스를 위한 음성인식 성능은 TV 자체에서 나오는 잡음 및 일반적인 생활환경에서 생성되는 잡음에 의해서 그 성능이 크게 나빠진다. 일반적인 생활환경에서 고려 할 수 있는 잡음은 정상성 잡음과 비정상성 잡음으로 구분된다. 정상성 잡음은 에어컨 작동소리, 컴퓨터 팬과 같은 잡음을 말하며, 비정상성 잡음은 TV 스피커 잡음, 생활 잡음과 같은 것이 있다.

일반적으로 정상성 잡음을 제거하는 방법으로는 단일채널을 기반으로 한다. 단일채널 잡음제거 알고리즘에서는 잡음의 통계량이 시간에 따라 거의 변하지 않는다는 가정에 기반을 하여 잡음 제거하며, 그 방법으로는 위너필터와 칼만 필터 방법이 있다[1-4]. 반면 비정상성 잡음제거 방법에는 다 채널 마이크로폰을 기반으로 한 빔포밍(beamforming), BSS (blind source separation) 알고리즘이 있다[5-8]

본 논문에서는 지능형 TV의 스피커에서 재생되는 비정상성 잡음의 제거를 주요 목표로 한다. TV의 스피커에서 재생되는 잡음은 전기적 결선을 통해서 그 신호를 미리 취득 할 수 있다는 특징을 가지고 있다. 이러한 조건에서는 기존의 LMS (least-mean square) 알고리즘을 활용할 수 있는데, 충분한 잡음이 제거되기 위해서 필터의 수렴시간이 필요하다는 것과 목표 음성 구간에서의 음성 왜곡 등의 문제점이 있다. 이러한 문제점을 해결하기 위해서 본 연구에서는 TV 스피커에서 취득한 참조 잡음 신호로부터 마이크로폰으로 수신되는 잡음 신호의 전력 스펙트럼을 추정한 후에 위너필터 기반의 잡음제거를 수행한다. 또한, 잡음제거 후에 잔존하는 잡음은 ETSI (European Telecommunications Standards Institute) AFX (advanced feature extraction) 에서 채택하고 있는 후처리를 통해서 추가적으로 제거한다[3].

Ⅱ. 관련 연구

2.1. LMS 알고리즘

LMS 알고리즘 동작을 위한 기본 개념도를 그림 1에 나타내었다.

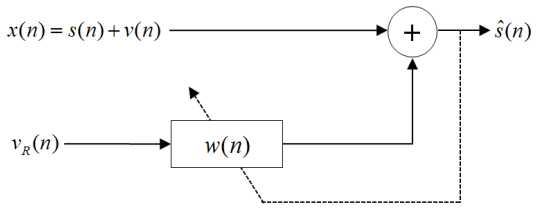

그림 1. LMS 알고리즘 기반 잡음제거
Fig. 1 LMS-based noise reduction

그림 1에서, $x(n)$은 입력신호, $s(n)$은 목표 음성신호, $v(n)$은 잡음 신호이다. $v_R(n)$은 참조 잡음 신호이고, $s(n)$과의 상호 상관도는 0으로 가정한다. $w(n)$은 필터 계수를 나타낸다. $\hat{s}(n)$은 잡음이 제거된 목표 신호이다. LMS 알고리즘에 의한 필터 계수의 갱신을 식 (1)에 나타내었다[4][9].

$$w_{n+1}(k) = w_n(k) + \mu x(n)\hat{s}(n-k) \quad (1)$$

식 (1)에서 $w_n(k)$는 n 번째 입력 샘플에 대한 k 번째 필터 계수의 값이다. μ는 적응 필터의 학습률이다. 보폭 변수가 클 때, 필터의 수렴은 빠를 수 있으나 잔존 잡음의 크기가 커지며 때에 따라서 필터가 발산할 수 있다[9]. 반면, 보폭 변수가 작을 때는 필터의 수렴이 늦어져서 충분한 잡음제거가 이루어지지 못하는 단점이 있다.

2.2. 위너필터

그림 2에서 위너필터 기반 잡음제거의 개념도를 나타내었다.

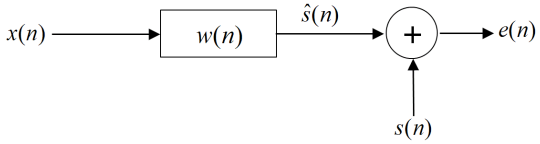

그림 2. 위너필터 기반 잡음제거
Fig. 2 Wiener filter-based noise reduction

위너필터 추정을 위한 비용 함수는 식 (2)와 같이 표현할 수 있다.

$$J = E[e^2(n)] = E[(\hat{s}(n) - s(n))^2] \tag{2}$$

목표 음성신호의 최적 추정치인 $\hat{s}(n)$은 식 (3)과 같이 표현될 수 있다.

$$\hat{s}(n) = \sum_{l=-\infty}^{\infty} w(l)x(n-l) \tag{3}$$

또한, 식 (2)의 비용함수는 $w(l)$에 대한 2차함수 이므로 최적점이 1개만 존재한다. 따라서, 식 (2)를 $w(l)$로 편미분 후에 0으로 놓으면 식 (4)를 얻을 수 있다.

$$\sum_{l=-\infty}^{\infty} w(l)r_x(m-l) = r_{sx}(m), \ \forall m \tag{4}$$

식 (4)에서 $r_x(m) = E[x(n)x(n-m)]$, $r_{sx}(m) = E[s(n)x(n-m)]$을 나타낸다. 또한, 목표 음성신호 $s(n)$과 가산성 잡음신호 $v(n)$이 WSS (wide-sense stationary)임과 동시에 통계적으로 독립이라고 가정하면 식 (5)를 얻을 수 있다.

$$\sum_{l=-\infty}^{\infty} w(l)(r_x(m-l) + r_v(m-l)) = r_d(m) \tag{5}$$

주파수 영역에서의 최적 위너필터의 계수를 추정하기 위하여 식 (5)의 양변에 DTFT (discrete-time Fourier transform)을 취하면 식 (6)을 얻을 수 있으며 그것의 시간 응답은 IDTFT (inverse DTFT)를 통하여 구할 수 있다.

$$W(e^{jw}) = \frac{P_s(e^{jw})}{P_s(e^{jw}) + P_v(e^{jw})} = \frac{SNR(e^{jw})}{1 + SNR(e^{jw})} \tag{6}$$

식 (6)에서 $P_s(w)$와 $P_v(w)$은 목표 음성신호와 잡음신호의 전력 스펙트럼을 나타낸다. 식 (6)에서 알 수 있듯이 주파수 영역에서 SNR을 알 수 있으면 최적 잡음제거를 위한 위너필터를 추정할 수 있다.

Ⅲ. 제안된 알고리즘

제안된 알고리즘의 목표는 지능형 TV 스피커를 통해 재생되는 잡음을 효과적으로 제거하는 것이다. 제 1장 서론에서 언급한바와 같이 LMS 알고리즘의 단점을 개선하기 위해서 본 연구에서는 위너필터 기반의 잡음제거를 구현하며 이를 위해서 주파수 영역에서의 SNR 추정에 연구의 초점을 맞춘다. 그림 3은 제안된 알고리즘의 전체적인 구조를 나타내고 있다. 총 2개의 입력을 사용하며 하나의 입력은 TV 스피커에 전기적 결선을 하여 재생되는 신호만을 취득한다. 또 다른 입력은 마이크로폰으로부터 신호가 수신되며 목표 음성신호와 TV 스피커로부터 재생되는 잡음신호가 가산되어 있다. $X(k)$과 $V^{(R)}(k)$는 $x(n)$와 $v^R(n)$의 DFT (discrete Fourier transform) 결과이며 k 는 이산 주파수 인덱스이다. A(k)는 TV 스피커와 마이크로폰 사이에 존재하는 경로 보상 함수, $\hat{P}_v(k)$는 마이크로폰을 통해 수신되는 잡음신호 전력 스펙트럼의 추정치이다. 그림 3의 각 블록에 대한 상세한 설명은 다음과 같다.

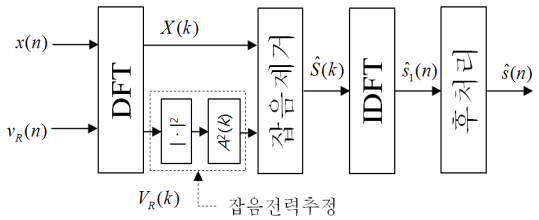

그림 3. 제안된 잡음제거 알고리즘의 구조
Fig. 3 Structure of proposed noise reduction

3.1. 잡음 전력 추정

마이크로폰으로 수신되는 TV 잡음의 전력 스펙트럼을 추정하기 위해서 TV 스피커와 마이크로폰 간에 존재하는 경로 전달함수를 추정한다. 추정 방법을 식 (7)에 나타내었다.

$$A(k)=\sqrt{\frac{\sum_{t=0}^{T-1}|N(t,k)|^2}{\sum_{t=0}^{T-1}|N_{ref}(t,k)|^2}} \tag{7}$$

식 (7)에서 $N(t,k)$는 마이크로폰에서 수신되는 TV 잡음 신호의 DFT 결과, $N_{ref}(t,k)$는 참조 잡음 채널에서 취득한 신호의 DFT 결과이다. t 는 주파수 분석을 위한 단구간 분석 프레임의 인덱스, T는 추정에 사용된 단구간 분석 프레임의 개수이다. 본 연구에서의 실험에서는 정확한 추정을 위하여 목표 음성 및 주변 잡음이 없는 조건 하에서 TV 스피커로부터 백색잡음을 재생하여 수식 (7)을 계산하였다.

3.2. 잡음제거

제안된 알고리즘에서 위너필터를 구현하려면 각 주파수 인덱스에 대해서 SNR을 추정해줘야 한다. 따라서, 식 (7)의 경로 전달함수를 이용하여 식 (8)의 목표 음성신호에 대한 전력 스펙트럼의 추정치를 구할 수 있다.

$$P_{\hat{s}}(k)=P_x(k)-A^2(k)P_{V_R}(k) \tag{8}$$

식 (8)에서 $P_x(k)$는 마이크로폰에 수신되는 입력신호의 전력 스펙트럼을 나타내며 $P_{V_R}(k)$는 참조 잡음 신호의 전력 스펙트럼을 나타낸다. 따라서, 각 주파수 인덱스에서의 SNR은 식 (9)와 같이 표현할 수 있다.

$$SNR(k)=\frac{P_{\hat{s}}(k)}{A^2(k)P_{V_R}(k)} \tag{9}$$

식 (9) 및 식 (6)을 통하여 추정한 위너필터의 주파수 응답 $W(k)$를 식 (10)과 같이 마이크로폰 입력 신호의 주파수 응답에 곱할 경우에 TV 잡음이 제거된 목표 음성의 주파수 응답을 구할 수 있다.

$$\hat{S}_1(k)=W(k)X(k) \tag{10}$$

최종적으로 IDFT (Inverse DFT)를 통하여 시간 영역에서 잡음이 제거된 신호인 $\hat{s}_1(n)$을 얻을 수 있다. 이 때, DFT 및 IDFT를 위한 분석 신호의 길이가 입력 단구간 신호의 길이보다 일반적으로 더 길기 때문에 중첩 가산 방식의 신호 합성과정을 수행한다[10].

3.3. 후처리

제안된 방식에 의한 잡음제거를 수행하더라도 잔존하는 잡음이 존재하므로 추가적인 단일채널기반의 잡음제거를 추가적으로 수행하면 잡음제거의 성능을 더욱 높일 수 있다. 본 연구에서 채택한 후처리 기법은 ETSI AFX에서 채택하고 있는 2 단계 위너필터 기법이다[3]. 이 방식에서는 배경 잡음의 시간축에서의 주파수 변화도가 크지 않다는 가정 하에 음성 검출과정을 통하여 잡음의 전력 스펙트럼을 지속적으로 추정한다. 그런 후에 검출된 음성 구간에서 3.2절의 과정을 통하여 잡음이 제거된 신호를 구하게 된다.

Ⅳ. 실험 및 결과

그림 4에서 본 연구의 실험 환경을 나타내었다. 실험을 위한 장비들은 5 m x 6 m x 3 m 크기의 일반 사무실 환경에 배치되었으며, 목표 음성원과 TV와의 거리는 3 m, TV의 크기는 49 인치였다.

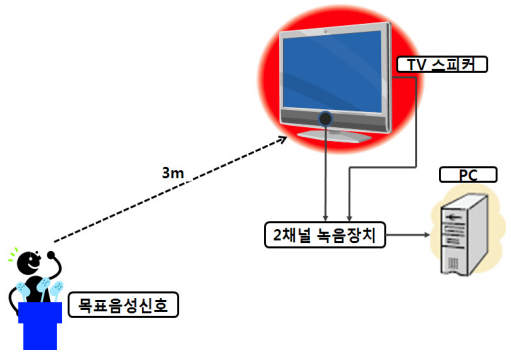

그림 4. 실험 환경
Fig. 4 Experiment environment

4.1. DB 구성

실험을 위해 사용된 모든 음성 데이터베이스는 스테레오 라인 입력이 가능한 외장형 사운드 카드를 통해서 수집되었다. 스테레오 입력중 하나는 TV의 스피커와 전기적 결선되며 다른 하나는 마이크로폰과 연결된다. 녹음된 DB의 표본화율은 모두 16 kHz이다. 목표 음성신호는 TV에 부착된 마이크로폰 3 m 정면에서 한국인명 2000개를 스피커를 통해 재생 시켜 녹음하였다. 지능형 TV에서 재생되는 잡음신호는 'West Life'의 곡 'Mandy'과 예능 프로그램의 대화 일부분을 사용하였다. 목표 음성신호와 지능형 TV에서 재생되는 잡음신호는 각각 별도로 녹음되며, 두 신호는 식 (11)에 의해서 인공적으로 가산하여 원하는 SNR을 맞춘다.

$$x(n) = s(n) + \alpha \times v(n) \tag{11}$$

여기서, $s(n)$은 목표 음성신호, $v(n)$은 TV 잡음, α는 잡음 증폭 인자이다. 본 연구에서 고려한 입력 SNR은 -5, 0, 5, 10 , 15, 20 dB 였다.

4.2. 실험 조건

제안한 알고리즘의 성능은 NLMS (normalized LMS) 알고리즘과 비교되었다. 제안한 알고리즘의 구현을 위해서 매 10 ms 단위로 30 ms 의 분석 프레임에 대해서 512-FFT (fast Fourier transform)을 수행하였다. 또한, 식 (7)의 경로 전달함수 추정을 위해서 백색잡음 10초를 TV 스피커를 통해서 재생하였다. NLMS 알고리즘의 구현을 위한 적응 FIR (finite impulse response) 필터의 길이는 127 이였으며 적응 필터의 학습률은 0.3으로 설정하였다.

4.3. 성능 평가

성능 평가를 위하여 입력 SNR에 따른 음성 인식률을 측정하였다. 음성 인식률 측정을 위하여 트라이폰 HMM (hidden Markov model)기반 음성 인식기를 이용하였다 [11]. HMM의 훈련을 위한 특징 파라미터로 MFCC (mel-frequency cepstral coefficient) 기반의 39차 벡터를 이용하는 잡음이 토크쇼 일 경우에 대한 음성 인식률을 나타내었다. 입력 SNR이 5 dB 일 때, 가공하지 않은 입력신호에 대해서는 64.3 %, NLMS만 수행 할 경우는 72.9 %, 제안된 방식을 수행하였을 경우에는 82.5 %, NLMS + 후처리 방식은 77.5 %, 제안된 방식 + 후처리 방법이 86.7 %의 음성인식률을 나타내었다. 그림 6은 TV 스피커 잡음이 음악 일 경우 음성 인식률을 나타내었다. 음성인식률의 추이는 음악잡음일 때와 유사하였다. 입력 SNR이 5 dB일 때, 가공하지 않은 입력신호에 대해서는 67.9 %, NLMS만 수행 할 경우는 77.7 %, 제안된 방식을 수행하였을 경우에는 86.1 %, NLMS + 후처리 방식은 85.8 %, 제안된 방식 + 후처리 방법이 88.0 %의 음성인식률을 나타내었다. 잡음의 종류에 상관없이 제안된 잡음제거 방식이 NLMS 기반 잡음제거 방식에 비해서 더 좋은 성능을 나타냄을 확인할 수 있다.

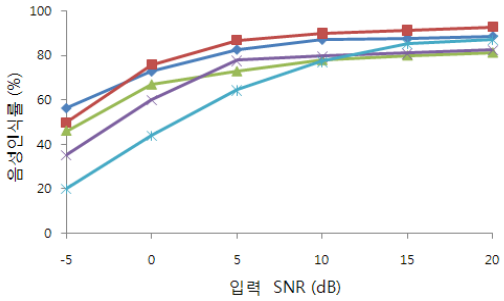

그림 5. TV 스피커 잡음이 토크쇼 일 경우의 음성 인식률('■' : 제안 된 알고리즘 + 후처리, '◆' : 제안 된 알고리즘, '▲' : NLMS, '×' : NLMS + 후처리, '*' : 잡음제거 전 입력)
Fig. 5 Speech recognition rates for a talkshow as TV speaker noise ('■' : proposed algorithm + postprocessing, '◆' : proposed algorithm, '▲' : NLMS, '×' : NLMS + postprocessing, '*' : noisy input)

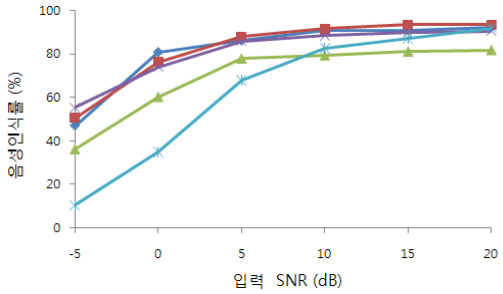

그림 6. TV 스피커 잡음이 음악일 경우의 음성 인식률
Fig. 6 Speech recognition rates for music as TV speaker noise

이러한 결과를 얻은 이유는 NLMS 알고리즘의 경우 적응 필터의 수렴에 필요한 시간 동안 잡음이 충분히 제거될 수 없다는 것과 목표 음성구간에서 표본 단위의 적응 필터 갱신이 이루어질 경우에 필연적으로 음성의 왜곡이 발생할 수밖에 없다는 것을 들 수 있다. 반면, 제안된 알고리즘은 단구간 분석 프레임에서 최적의 잡음제거를 위한 위너필터를 추정할 수 있다는 것과 상대적으로 NLMS에 비해서 목표 음성구간의 왜곡을 감소시킬 수 있다는 장점 때문에 더 좋은 성능을 나타낼 수 있었다.

V. 결 론

본 논문에서는 지능형 TV의 음성인식을 위한 참조잡음 신호 기반 음성개선 알고리즘을 제안하였다. 이를 위하여 먼저, TV 스피커와 마이크로폰 간의 경로 함수를 추정하였으며 그것을 활용하여 최적 잡음제거를 위한 위너필터를 추정하였다. 추가적으로 잔존하는 배경 잡음을 제거하기 위하여 ETSI AFX에서 채택하고 있는 2단계 위너필터를 후처리 알고리즘에서 채택하였다. 제안한 잡음제거 알고리즘은 기존의 NLMS 기반의 알고리즘에 비해서 모든 입력 SNR에서 더 우수한 성능을 보였다. 향후 연구로는 참조 잡음신호를 얻기 위한 전기적 결선 문제를 해결하기 위해서 빔포밍 기술을 활용할 예정이다. 즉, TV 스피커 방향으로의 빔포밍을 수행한다면 어느 정도의 추정이 가능할 것으로 판단하며 부수적으로 TV 주변의 비정상성 생활 잡음도 제거할 수 있을 것으로 사료된다.

참고문헌

[1] S. Jeong and M. Hahn, "Speech quality and recognition rate improvement in car noise environments," *Electronics Letters*, vol. 37, no. 12, pp. 801-802, 2001.

[2] Y. Ephraim, and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 32, no. 6, pp. 1109-1211, 1984.

[3] ES 202 212 V1.1.2., "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; extended advanced front-end feature extraction algorithm; compression algorithm; back-end speech reconstruction algorithm," *ETSI Standard*, 2005.

[4] B. Widrow, and S. D. Stearns, *Adaptive Signal Processing*, Prentice Hall, 1985.

[5] 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, 1998.

[6] M. Brandstein, and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, Springer, 2001.

[7] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing* (*Springer Topics in Signal Processing*), Springer, 2008.

[8] O. Hoshuyama, et al., "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. Signal Proc.*, vol. 47, no. 10, pp. 2677-2688, 1999.

[9] S. Alexander, *Adaptive signal Processing: Theory and Applications*, Springer-Verlag, 1986.

[10] A. Oppenheim, R. Schafer, *Discrete-time signal processing*, Pearson, 2007.

[11] L. Rabiner, and B. H. Juang, *Fundamentals of speech recognition*, Prentice Hall, 1996.

저자소개

정상배(Sangbae Jeong)

1997년 3월 부산대학교
전자공학과 졸업(공학사)
1999년 2월 한국과학기술원 전기
및 전자공학과 졸업
(공학석사)
2002년 8월 한국정보통신대학교 공학부 졸업
(공학박사)
2002년 9월~2006년 2월 삼성종합기술원 책임연구원
2006년 3월~2009년 2월 한국정보통신대학교
디지털미디어연구소(연구조교수)
2009년 3월~현재 경상대학교 전자공학과/
공학연구원(조교수)
※관심분야 : 음성인식, 음성/오디오부호화