비음수 행렬 인수분해 기반의 음성검출 알고리즘

정회원 강 상 익*, 종신회원 장 준 혁**

Voice Activity Detection Based on Non-negative Matrix Factorization

Sang-Ick Kang* *Regular Member*, Joon-Hyuk Chang** *Lifelong Member*

요 약

본 논문에서는 비음수 행렬 인수분해 기법을 기반으로 한 새로운 음성 검출 (Voice Activity Detection, VAD) 알고리즘을 제안한다. 먼저, 기존의 통계모델기반의 음성검출기를 분석하고, 이를 기반으로 비음수 행렬 인수분해를 통해 도출한 입력 기초 벡터와 잡음 기초 벡터 차이로 음성의 유무를 판단한다. 이때 최적의 문턱값을 찾기 위해 통계모델 기반의 음성검출기에 의해 추정된 잡음 구간에서 NMF 결과의 분포에 따라 최적화된 문턱값을 비음수 행렬기반의 음성 검출 알고리즘에 적용하는 방법을 제안한다. 실험 결과 기존의 통계적 모델 기반의 음성검출기에 비해 6.75%의 성능향상을 가져왔다.

Key Words : Voice Activity Detection, Non-negative Matrix Factorization

ABSTRACT

In this paper, we apply a likelihood ratio test (LRT) to a non-negative matrix factorization (NMF) based voice activity detection (VAD) to find optimal threshold. In our approach, the NMF based VAD is expressed as Euclidean distance between noise basis vector and input basis vector which are extracted through NMF. The optimal threshold each of noise environments depend on NMF results distribution in noise region which is estimated statistical model-based VAD. According to the experimental results, the proposed approach is found to be effective for statistical model-based VAD using LRT.

Ⅰ. 서 론

음성과 비음성 구간을 검출하는 음성 검출기 (voice activity detector, VAD)는 음성 부호화, 음성인식 그리고 음향학적 반향제거기 등 음성 통신 시스템에서 많이 적용된다. 특히, 음성 통신 시스템의 대역폭을 효율적으로 사용하기 위해서 필수적으로 요구되며 Ephraim과 Malah의 연구에서 시작된 minimum mean square error (MMSE) 기반의 음성 향상 기법에 사용된 음성의 존재와 부재에 대한 통계적 모델을 음성 검출기 (likelihood ratio test, LRT) 에 적용한 것이 우수한 음성 검출 성능을 가진 것으로 알려져 있다[1-8].

기존의 음성 검출 알고리즘에서는 음성과 잡음에 대한 통계적 모델을 기반으로 음성의 활성 여부를 판단하기 때문에, 신호 대 잡음비가 높은 신호에서는 비교적 정확한 음성 검출이 가능하지만 상대적으로 열악한 잡음 환경에서는 음성 검출의 성능이 급격히 저하되는 단점이 있다.

비음수 행렬 인수분해는 이미지 등의 패턴 학습과 패턴 인식에 우수한 성능을 보이는 알고리즘이다. 특

※ 본 연구는 지식경제부 및 한국산업기술평가관리원의 IT핵심기술개발사업의 일환으로 수행하였음. [고성능 가상머신 규격 및 기술 개발, 2009-S-036- 01] 및 이 논문은 2009년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(2009-0085162)

* 인하대학교 전자공학과 DSP연구실(sikang@dsp.inha.ac.kr), ** 인하대학교 전자공학과(changjh@inha.ac.kr)

논문번호 : KICS2010-05-213, 접수일자 : 2010년 5월 14일, 최종논문접수일자 : 2010년 7월 13일

히 다수의 입력 데이터에서 최적의 기초 패턴을 분리하여 이들의 선형 조합으로 전체 데이터를 근사할 수 있기 때문에 데이터 특징 추출에도 유용하다[9-10].

본 논문에서는 비음수 행렬 인수분해 기반의 새로운 음성 검출 기법을 제안하였다. 그리고 잡음 신호로부터 추출한 기초 벡터들과 입력 신호와의 오차를 계산하였고, 이를 기반으로 음성의 활성 구간을 구분하였다. 이 때 최적화된 문턱값을 선택하기 위해 잡음 추정 구간에서의 오차값 분포에 따라 잡음 환경을 추정하여 문턱값을 선정하였다. 실험 결과 제안된 음성 검출 알고리즘이 기존의 통계적 모델 기반 음성 검출기에 비해 더 우수한 성능을 보였다.

본 논문의 II 장에서는 통계적 모델 기반의 음성검출기, III 장에서는 비음수 행렬인수분해 기반의 음성검출기에 대해 살펴본다. 그리고 IV 장에서는 실험결과를 종합적으로 검토하고 V 장에서 결론을 맺는다.

Ⅱ. 통계모델 기반의 음성 검출기의 이해

시간축 상에서 원래의 음성신호 $x(t)$에 잡음신호 $n(t)$이 인가된 입력신호 $y(t)$을 discrete Fourier transform (DFT)을 통해 주파수 축으로 변환되어 아래와 같이 표현된다.

$$\boldsymbol{Y}(t) = \boldsymbol{X}(t) + \boldsymbol{N}(t) \tag{1}$$

여기서 $\boldsymbol{Y}(t) = [Y_1, Y_2, ..., Y_M]$, $\boldsymbol{X}(t) = [X_1, X_2, ..., X_M]$, 그리고 $\boldsymbol{N}(t) = [N_1, N_2, ..., N_M]$는 각각 잡음에 오염된 음성신호, 원래의 음성신호, 잡음신호의 DFT 계수 벡터를 나타낸다. 주어진 가설 H_0, H_1이 각각 음성의 부재와 존재를 표현한다고 하면 각 주파수 채널별로 다음과 같이 기술된다.

$$H_0 : \text{speech absent} \quad : \quad Y_k(t) = N_k(t) \tag{2}$$

$$H_1 : \text{speech present} : \quad Y_k(t) = X_k(t) + N_k(t). \tag{3}$$

음성과 잡음신호의 스펙트럼이 복소 가우시안 분포를 따른다는 가정으로부터 가설 H_0와 H_1을 조건으로 한 확률밀도함수는 아래와 같이 주어진다[3].

$$p(Y_k|H_0) = \frac{1}{\pi\lambda_{d,k}} \exp\left\{ -\frac{|Y_k|^2}{\lambda_{d,k}} \right\} \tag{4}$$

$$p(Y_k|H_1) = \frac{1}{\pi[\lambda_{d,k} + \lambda_{x,k}]} \exp\left\{ -\frac{|Y_k|^2}{\lambda_{d,k} + \lambda_{x,k}} \right\} \tag{5}$$

여기서 $\lambda_{x,k}$와 $\lambda_{d,k}$는 각각 채널별 음성과 잡음의 분산이며, 이 때 k번째 주파수 밴드에 대한 우도비는 아래와 같이 구한다.

$$\Lambda_k \equiv \frac{p(Y_k|H_1)}{p(Y_k|H_0)} = \frac{1}{1+\xi_k} \exp\left\{ \frac{\gamma_k \xi_k}{1+\xi_k} \right\} \tag{6}$$

여기서 $\xi_k = \lambda_{x,k}/\lambda_{d,k}$와 $\gamma_k = Y_k/\lambda_{d,k}$는 각각 a priori signal-to-noise ratio (SNR)와 a posteriori SNR이다[3]. 음성 부재 구간에서 갱신되는 잡음 신호로부터 구한 잡음 분산 $\lambda_{d,k}$를 이용하여 a posteriori SNR γ_k를 추정하며, 또한 a priori SNR ξ_k는 decision-directed (DD) 방식을 이용하여 아래와 같이 추정한다[1].

$$\hat{\xi}_k(t) = \alpha \frac{|\hat{X}_k(t-1)|^2}{\lambda_{d,k}(t-1)} + (1-\alpha) P[\gamma_k(t) - 1] \tag{7}$$

여기서 $|\hat{X}_k(t-1)|^2$은 이전 프레임에서 추정된 음성 신호의 k번째 스펙트럼 성분의 크기에 대한 추정치이며, MMSE에 기반하여 구한다[3]. 또한 α는 가중치 값이며, 연산자 $P[\bullet]$은 아래와 같이 정의된다.

$$P[x] = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

기존의 일반적인 통계적 모델 기반의 음성 검출기에 대한 결정식은 각각의 주파수 채널에서 구해진 우도비를 기하 평균하여 아래와 같이 음성 검출 여부를 판단한다[2-8].

$$\log \Lambda(t) = \frac{1}{M} \sum_{k=1}^{M} \log \Lambda_k \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{9}$$

여기서 M은 전체 주파수 대역의 개수이며, η는 음성 검출 문턱값이다.

Ⅲ. Non-negative Matrix Factorization (NMF)

3.1 비음수 행렬 인수분해

비음수 행렬 인수분해 (Non-negative Matrix Factorization, NMF)는 Principal Components Analysis (PCA), Vector Quantization (VQ) 와 마찬가지로 기

초 벡터들의 선형 결합으로 근사하여 행렬을 분해하는 기법으로서 NMF는 특별히 모든 성분이 비음수인 제약을 가진다[9-10].

비음수 성분으로 구성된 $n\times m$ 행렬 V는 $V\approx WH$ 형태, 즉

$$V_{i\mu}\approx (WH)_{i\mu}=\sum_{a=1}^{r}W_{ia}H_{a\mu} \tag{10}$$

과 같이 인수분해 된다. 여기서 행렬 V 는 $n\times 1$ 벡터가 m 개 결합된 것으로 볼 수 있으며, 행렬 W 는 $n\times r$ 크기로 역시 r 개의 $n\times 1$ 벡터 집합, H 는 $r\times m$ 크기의 계수 행렬이 된다. 즉 m 개의 n 차 데이터 벡터들을 r 개의 n 차 기초 벡터들의 선형 결합으로 표현된다.

일반적으로 행렬의 비음수 행렬 인수분해 과정은 행렬 W와 H을 반복적으로 갱신하여 행렬 V 와 WH 간의 거리를 최소화 하는 방향으로 근사한다. 따라서 어떠한 거리 함수 혹은 목적 함수를 적용할 것인가에 따라 성분 값 갱신 식이 달라진다. 일반적으로 널리 사용되는 목적 함수로는 Euclidean distance $\|V-WH\|^2=\sum_{ij}(V_{ij}-\sum_{a=1}^{r}W_{ia}H_{aj})^2$이며 이것을 최소화 하는 성분 값 갱신식은

$$H_{\alpha\mu}\leftarrow H_{\alpha\mu}\frac{(W^TV)_{\alpha\mu}}{(W^TWH)_{\alpha\mu}} \tag{11}$$

$$W_{i\alpha}\leftarrow W_{i\alpha}\frac{(VH^T)_{i\alpha}}{(WHH^T)_{i\alpha}} \tag{12}$$

과 같다[10].

3.2 비음수 행렬 인수분해 기반의 음성 검출 기법

비음수 행렬 인수분해 기법은 데이터 벡터를 기초 벡터들의 선형 결합으로 근사한다. 따라서 사전에 잡음 신호로부터 기초 벡터를 추출하여 입력 신호의 기초벡터와의 거리를 계산하여 얻은 값으로 입력 신호와 잡음 신호의 유사도를 판단할 수 있다.

본 논문에서는 2장에서 잡음 분산 $\lambda_{d,k}$를 이용하여 추정한 *a posteriori* SNR와 DD 방식을 이용하여 추정한 *a priori* SNR을 특징 벡터로 두고, 시작 부분의 $m+9$ 개 프레임을 음성 부재 구간으로 가정한 후, m 개 프레임의 슈퍼프레임들을 비음수 행렬 인수분해 하여 배경 잡음 신호의 기초 벡터 W 를 추출하고 1 프레임씩 이동하여 기초 벡터를 추출한다. 이 때 m값을 5로 사용하였다. 총 10 개의 기초 벡터의 평균 $\overline{W}$와 입력 신호의 기초 벡터들 사이의 거리를 계산하여 음성 활성 여부를 다음과 같이 판단한다.

$$\|\overline{W}-W\|^2 \underset{H_0}{\overset{H_1}{\gtrless}} \eta' \tag{13}$$

여기서 η'은 NMF 기반의 음성 검출기에 적용되는 문턱값이다.

추출된 기초 벡터 $\overline{W}$는 배경 잡음 신호로부터 추출된 것이므로 이 기초 벡터들과 유사한 신호가 입력될 경우 거리는 상대적으로 작으며 잡음 신호와 특성이 다른 음성 신호와의 거리는 상대적으로 매우 크다. 그림 1은 white 잡음환경에서 잡음 신호의 기초벡터 $\overline{W}$와 입력 신호의 기초벡터 W 사이의 거리를 보여주고 있다.

그림 2는 NMF 기반의 음성 검출기에서 문턱값 η'의 변화에 따른 음성 검출 확률 P_e의 변화 곡선이다. 그림 2와 같이 NMF 기반의 음성 검출 결과가 잡음환경에 따라 각각의 최적화된 문턱값을 가지는 것을 알 수 있으며 따라서 최적화된 성능을 위하여 잡음환경에 따른 문턱값이 존재해야 한다.

본 논문에서는 NMF 기반의 음성 검출기에서 최적화 된 문턱값을 적용하기 위하여 통계적 모델 기반의 음성검출기를 도입한다. NMF 결과를 사용하여 잡음환경을 인식하여 잡음 환경에 해당하는 최적 문턱값을 적용하기 위해서는 음성 구간을 최대한 배제하는

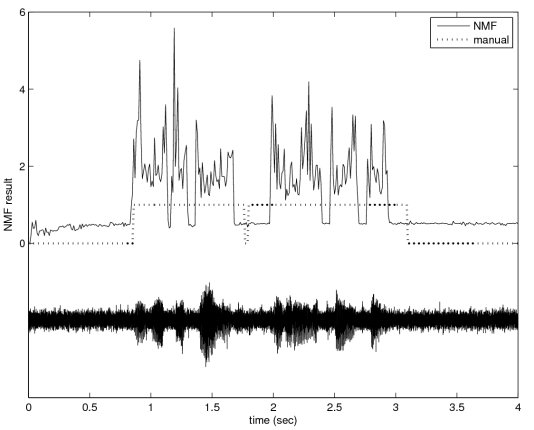

그림 1. White 잡음 5 dB SNR에서 NMF 음성검출기 출력 결과

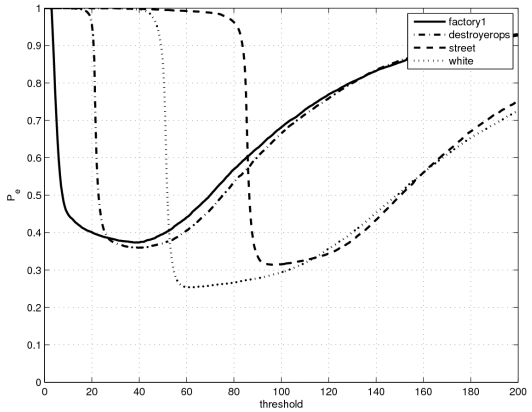

그림 2. 문턱값에 따른 P_e 값의 변화 (SNR =5 dB)

것이 바람직하다. 잡음 구간을 효과적으로 추정하기 위해 통계적 모델 기반의 음성검출기를 적용하여 비음성일 경우 NMF 결과 분포를 살펴본다. 그림 3은 통계적 모델 기반의 음성검출기를 통해 잡음 환경이라고 추정된 프레임에서 NMF 결과 분포를 보여준다. 그림과 같이 NMF 결과 분포는 각각의 잡음 환경에 따라 명확하게 구분할 수 있다. 따라서 각각의 잡음 환경에 대해 최적화된 문턱값을 아래와 같이 적용한다.

$$\Xi = \frac{1}{b}\sum_{t=1}^{50} \| \overline{W}(t) - W(t) \|^2 \Big|_{\log\Lambda(t) < \eta} \tag{14}$$

$$\eta' = \begin{cases} \eta'_1 & \Xi < \Xi_0 \\ \eta'_2 & \Xi_0 \le \Xi < \Xi_1 \\ \eta'_3 & \Xi_1 \le \Xi < \Xi_2 \\ \eta'_4 & \Xi_2 \le \Xi \end{cases} \tag{15}$$

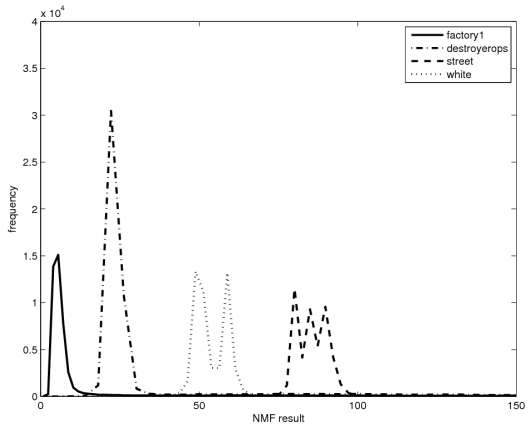

그림 3. 잡음 추정 구간에서 NMF 결과값 분포

여기서 b는 $\log\Lambda(t) < \eta$ 인 프레임의 개수를 나타낸다. 도출된 최적화된 문턱값 η'를 식 (13)에 적용하여 최종적으로 음성을 검출한다.

주목해야 할 점은 통계적 모델기반의 음성검출 알고리즘을 적용하여 비음성 구간에서 NMF 분포에 따라 잡음환경에 최적화된 문턱값을 적용함으로써 보다 향상된 음성 검출 결과를 얻을 수 있었으며 단순히 음성만을 검출하는 것이 아니라 특정 잡음 환경을 검출하여 후처리하는 방법으로 응용할 수 있다.

Ⅳ. 실험 결과

본 논문에서 제안된 음성 검출 알고리즘의 성능을 평가하기 위하여 오경보 (false alarm) 와 누락 (missing) 을 포함한 음성 검출 오류 확률 P_e (speech detection error probability) 을 측정하였다. 실험에 사용된 데이터는 총 230초의 깨끗한 음성 데이터에 음성과 비음성 부분을 10 ms마다 수동으로 표시하였다. 분류된 음성 데이터의 음성 구간은 총 57.1%로 유성음 44.0%, 무성음 13.1%로 구성되었으며 잡음 환경을 만들기 위해 NOISEX-92 데이터베이스로부터 white, factory1, destroyer ops 잡음을 ETRI 데이터베이스로부터 street 잡음을 사용하였으며 각각 0, 5, 10, 15 dB SNR 으로 원래의 음성신호에 더하여 사용하였다. NMF에서 기초 벡터를 구하기 위해 사용한 특징벡터는 NMF가 가지고 있는 비음수라는 제약을 만족하고 기존의 음성 검출기에서 우도비를 도출하기 위해 사용되는 *a priori* SNR ξ_k, *a posteriori* SNR γ_k 총 32차를 사용하였으며 $m=5$, $r=3$으로 사용하였다. 잡음 기초 벡터는 1 프레임씩 이동하면서 총 10개의 기초 벡터의 평균으로 사용하였으며 이 때 기초 벡터는 업데이트 되지 않는다. 각각의 잡음을 추정하기 위하여 50 프레임에서의 NMF 결과 분포를 사용하여 분포의 평균값으로 잡음 환경을 구분하여 문턱값을 결정하였다. 추가적으로 시스템에서 학습되지 않은 대표적인 non-stationary 잡음 환경인 babble 잡음환경에서도 실험을 하였다.

음성 검출 실험 결과는 표 1에 나타나 있으며 NMF 기반의 음성 검출기와 통계적 모델 기반의 음성 검출기의 오경보와 누락을 포함한 음성 검출 오류 확률 P_e를 보여준다. 실험 결과를 분석해 보면 낮은 SNR에서 NMF 기반의 음성검출기에 최적화된 문턱값을 적용한 실험 결과 두드러진 성능향상을 보였다. 이것은 NMF 기반의 음성인식기가 통계적 기반의 음성인식

표 1. 다양한 잡음환경과 SNR에서 제안된 NMF 기반 음성검출기와 통계모델기반 음성검출기 P_e 성능 비교

Noise	Methods	SNR			
		0 dB	5 dB	10 dB	15 dB
street	LRT	40.38	38.28	29.45	24.33
	NMF	33.06	31.42	27.04	23.68
white	LRT	80.76	35.38	30.85	27.48
	NMF	55.84	25.66	23.30	21.65
factory1	LRT	64.11	45.84	33.74	23.33
	NMF	48.81	37.96	29.87	22.47
Destroyer ops	LRT	52.88	39.37	28.25	19.28
	NMF	44.67	36.16	26.01	17.96
Babble	LRT	68.30	51.23	38.68	25.18
	NMF	68.21	49.16	37.45	25.12

기보다 신호 대 잡음비의 영향을 덜 받는 것으로 볼 수 있으며 평균적으로 6.75%의 음성 검출 성능 향상을 확인할 수 있다.

V. 결 론

본 논문에서는 비음수 행렬 인수분해 기법을 적용한 음성 검출 알고리즘을 제안하였다. 비음수 행렬 인수분해 기법을 이용하여 도출한 잡음 신호 기초 벡터와 입력 신호 기초 벡터 사이의 오차를 기반으로 음성을 검출하는데, 이 때 우도비에 의해 추정된 음성부재 구간에서의 오차값 분포에 따른 최적화된 문턱값을 적용하는 음성 검출 방법을 제시하였다. 실험 결과를 통하여 제안된 음성검출 알고리즘이 기존의 통계 모델 기반 음성 검출기에 비해 성능이 우수함을 확인하였다.

참 고 문 헌

〔1〕 Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,” *IEEE Trans. Acoustics, Speech, Sig. Process.*, Vol.ASSP-32, No.6, pp.1190-1121, Dec. 1984.

〔2〕 J. Sohn and W. Sung, “A voice activity detector employing soft decision based noise spectrum adaptation,” Proc. *Int. Conf. Acoustics, Speech, and Sig. Process.*, Vol.1, pp. 365-368, May 1998.

〔3〕 J. Sohn, N. S. Kim, and W. Sung, “A statistical model-based voice activity detection,” *IEEE Sig. Process. Lett.*, Vol.6, No.1, pp.1-3, Jan. 1999.

〔4〕 Y. D. Cho and A. Kondoz, “Analysis and improvement of a statistical model-based voice activity detector,” *IEEE Sig. Process. Lett.*, Vol.8, No.10, pp.276-278, Oct. 2001.

〔5〕 J. -H. Chang, J. W. Shin, and N. S. Kim, “Voice activity detector employing generalised Gaussian distribution,” *Electron. Lett.*, Vol.40, No.24, pp.1561-1563, Nov. 2004.

〔6〕 J. -H. Chang, N. S. Kim, and S. K. Mitra, “Voice activity detection based on multiple statistical models,” *IEEE Trans. Sig. Process.*, Vol.54, No.6, pp.1965-1976, June 2006.

〔7〕 Y. C. Lee and S. S. Ahn, “Statistical model-based VAD algorithm with wavelet Transform,” *IEICE Trans. Fundamentals*, Vol.E89-A, No.6, pp.1594-1600, June 2006.

〔8〕 J. Ramirez, J. M. Gorriz, J. C. Segura, C. G. Puntonet, and A. J. Rubio, “Speech / non-speech discrimination based on contextual information integrated bispectrum LRT,” *IEEE Sig. Process. Lett.*, Vol.13, No.8, pp.497-500, Aug. 2006.

〔9〕 D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, Vol.401, pp.788-791, Oct. 1999.

〔10〕 D. D. Lee and H. S. Seung, “Algorithms for Non-negative Matrix Factorization,” *In Advances in Neural Information Processing Systems*, Vol.13, pp.556 – 562, 2001.

〔11〕 A. Varga and H. J. M. Steeneken, “Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” *Speech Communication*, Vol.12, No.3, pp.247-251, 1993.

강 상 익 (Sang-Ick Kang) 정회원

2007년 2월 인하대학교 전자공학과 학사

2009년 2월 인하대학교 전자공학부 석사

2009년 3월~현재 인하대학교 전자공학부 박사과정

<관심분야> 음성신호처리

장 준 혁 (Joon-Hyuk Chang) 종신회원

1998년 2월 경북대학교 전자공학과 학사

2000년 2월 서울대학교 전기공학부 석사

2004년 2월 서울대학교 전기컴퓨터공학부 박사

2000년 3월~2005년 4월 ㈜넷더스 연구소장

2004년 5월~2005년 4월 캘리포니아 주립대학, 산타바바라 (UCSB) 박사후연구원

2005년 5월~2005년 8월 한국과학기술연구원 (KIST) 연구원

2005년 9월~현재 인하대학교 전자공학부 조교수

<관심분야> 음성신호처리, 오디오신호처리, 통신신호처리, 휴먼/컴퓨터 인터페이스