기계학습을 이용한 음성 신호처리 연구동향

김 태 수
한국퀄컴 연구소

Ⅰ. 서 론

지난 3월 많은 사람들은 구글의 인공지능 프로그램 알파고(AlphaGo)가 바둑계 최강자 이세돌 9단에게 4대1로 승리는 거두는 것을 지켜보았다. 바둑은 4000년 역사를 자랑하는 동양문화의 정수로서 체스와는 달리 경우의 수가 우주의 원자수 보다 많다. 또한 전략이 훨씬 복잡하고 다양해 계산보다는 직관, 상황판단 등 인간의 고유 능력이 중요해 기계가 인간을 이길수는 없을 것이라고 생각되어왔다 이러한 바둑에서 인간이 기계에게 패배하는 모습을 지켜본 사람들이 받은 충격은 알파고 충격이라는 말로 표현되었다[1].

알파고가 기존의 인공지능 바둑프로그램과 다른 점은 바로 기계학습(Machine Learning) 분야에서 최근 비약적 발전을 이루어 내며 많은 응용분야를 만들어 내고 있는 딥러닝(Deep Learning) 기술을 이용하여 인간의 직관과 상황판단 능력을 흉내냈다는 점이다. 기계학습은 인공지능의 한 분야로 컴퓨터가 주어진 데이터를 통해서 스스로 학습하도록 하는 알고리즘을 연구하는 분야이다. 이는 문제에 대한 구체적인 지식을 통해서 사람이 직접 알고리즘을 고안해내 문제를 해결하는 것이 아니라 많은 데이터를 통해서 알고리즘이 그 문제에 가장 적절한 해결책을 찾아내도록 하는 것이다. 알파고는 기존 대국의 기보 데이터를 통해서 바둑의 판세를 읽고 어떤 수를 놓는 것이 유리한지 알고리즘이 스스로 학습하였다. 그리고 학습된 알고리즘끼리 서로 대국하여 더 많은 데이터를 확보하고 이를 통해 알고리즘의 성능을 더욱 향상시키는 방법을 사용하였다[2].

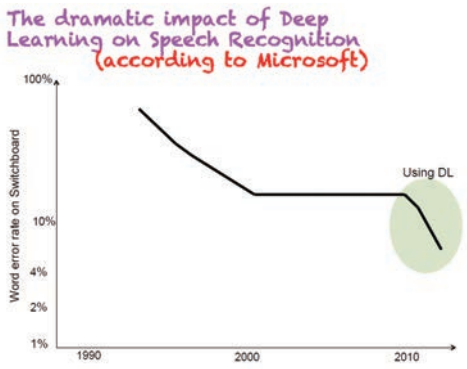
신호처리 분야에서도 기계학습을 이용한 연구들이 꾸준이 진행되어 왔다. 신호처리를 위한 기계학습 (Machine Learning for Signal Processing) 학회는 올해로 26년째 개최되고 있으며, 신호처리 분야의

최대 학회인 ICASSP에서는 매년 기계학습 관련 논문이 늘어나고 있다. 본 논고에서는 이 중 음성 신호처리의 대표적인 분야인 음성인식(speech recognition) 과 음성향상(speech enhancement)에서 기계학습이 쓰이고 있는 동향을 살펴보고자 한다.

Ⅱ. 관련 분야 동향

1. 음성인식

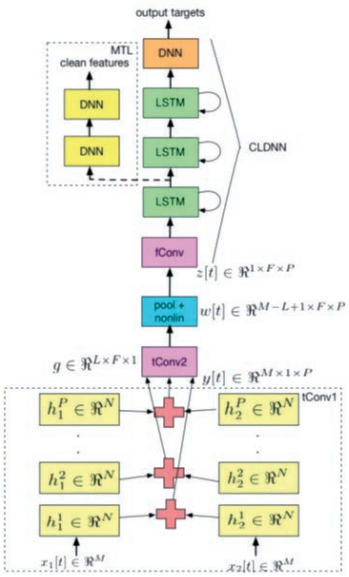
음성인식은 사람의 발성을 마이크를 통해 입력받아 문자로 바꿔주거나 해당 명령을 수행하도록 하는 기술이다. 1952년 미국의 벨 연구소(Bell Laboratories)에서 단일 숫자음성 인식 시스템 오드레이(Audrey) 개발을 통해서 시작되어 1963년 IBM은 세계 최초로 음성을 통해 16개의 영어단어를 인식할 수 있고 간단한 숫자 계산이 가능한 슈박스(Shoebox)라는 장비를 공개하였다. 이 후 많은 민간 및 정부 연구소들에서 연구를 진행하여 1980년대에는 인식할 수 있는 단어수가 1,000단어에서 1만 단어까지 늘어났으며, 1990년대에 이르러 음성인식이 상용화 될 수 있었다. 초창기 음성인식기에는 동적시간 정합(Dynamic Time Warping) 기술과 같이 입력된 음성 신호를 정해진 패턴과 매칭하는 방식인 템플릿 매칭기반의 알고리즘들이 사용되었으나 이 후 기계학습의 일종인 신경망(Neural Networks)이 사용되기도 하였다. 하지만, 데이터를 통한 학습이 오래 걸리고, 주어진 데이터에만 과도하게 잘 맞춰지고 실제 환경에서는 성능이 떨어지는 이른바 과적응(overfitting) 문제등으로 인해 연구자들의 관심에서 멀어졌다. 그 후 비슷한 시기에 제안된 은닉 마코프 모델(Hidden Markov Model)을 이용한 방법이 비교적 근래까지 주류가 되었다. 은닉 마코프 모델을 이용한 방법은 통계학적 방법으로서 주어진 데이터를 이용하여 모델의 파라메터들을 학습하는 기계학습의 일종이긴 하나 알고리즘을 최적화 하는데 음성학적 지식도 많이 사용되는 방법이기도 하다. 이 후 한동안 음성인식분야에 있어서 큰 진전이 없다가 2010년 마이크로소프트의 연구진은 은닉 마코프 모델의 관측확률(observation probability)을 심층 신경망(Deep Neural Networks)으로 만들어서 학습하는 이른바 딥러닝을 통해서 극적인 성능향상을 보여주게 되었다[4–5].



〈그림 1〉 연도별 음성인식 오인식률 추이 (출처 NIPS2015 tutorial)

〈그림 1〉은 마이크로소프트의 연도별 음성인식 오인식율 추이를 보여주는 그래프이다[3]. 이에 따르면 2000년까지 음성인식기의 성능은 꾸준히 개선이 되어 20% 정도의 단어 오인식률을 보여준다. 그리고 2000년 이후에는 다양한 연구가 여전히 진행되었지만 이렇다 할 성능 개선을 보여주지 못하고 비슷한 성능을 보여주다가 2010년 이후로 급격한 성능 향상을 이루게 되는데 이 때 적용된 방법들이 딥러닝에 기반한 알고리즘들이다. 최근에는 관측 확률만을 신경망을 통해서 학습하는 것이 아니라 은닉 마코프 모델까지도 순환 신경망(Recurrent Neural Networks)으로 대체되고 있다. 구글은 최근 순환 신경망을 이용하여 그들의 2000시간의 음성검색 데이터베이스에서 오인식률을 기존의 심층 신경망 모델에서 얻은 10.1%에서 8.5%까지 향상시켰다[6]. 바이두(Baidu)에서는 합성곱 신경망(Convolutional Neural Networks)과 순환 신경망을 같이 사용하여 중국어 검색어에서 사람수준의 인식성능을 얻었다고 보고하였다[7]. IBM에서도 합성곱 신경망과 순환 신경망을 같이 사용하여 2000시간의 대화체 음성 데이터베이스 (Switchboard Hub5-2000)에서 8%의 오인

최근에는 관측 확률만을 신경망을 통해서 학습하는 것이 아니라 은닉 마코프 모델 까지도 순환 신경망 (Recurrent Neural Networks)으로 대체 되고 있다.



〈그림 2〉 2개의 마이크 신호를 입력으로 하는 음성인식 신경망

식률을 얻었다[8].

최근의 신경망을 이용한 음성인식에서 주목할 점은 기존의 방법들이 음성인식 전단(front-end)인 특징추출 방법을 MFCC(Mel-Filterbank Cepstral Coefficient)와 같은 연구자들이 직접 고안한 알고리즘을 사용하지 않고 스펙트럼 또는 심지어 마이크 입력 신호를 그대로 사용하기도 한다는 점이다. 이렇게 입력된 신호에서 기계학습 알고리즘이 특징을 찾고 음성인식까지 한번에 하는 구조인 것이다. 이미 앞에서 언급한 최근 연구들에서는 스펙트럼을 입력으로 받아서 기존의 특징추출방법에 비해서 나은 성능을 보여주고 있다. 최근 구글에서는 마이크 신호를 입력으로 하는 음성인식 신경망을 발표하였다[9]. 〈그림 2〉는 2개의 마이크 신호를 입력으로 하는 음성인식 신경망의 구조이다. tConv1은 합성곱 신경망으로 2개의 마이크에서 공간 필터링을 하기위한 정보를 학습하는데 일종의 빔포밍 역할을 하도록 학습된다. tConv2는 또 다른 합성곱 신경망으로 이는 스펙트럼 분석의 역할을 하도록 학습된다. 이후 fConv에서는 스펙트럼에서 음성인식에 적합한 특징을 추출하는 역할을 하고 장단기 기억 순환 신경망층인 LSTM과 심층 신경망 층인 DNN으로 구성되어 최종 음성인식 결과를 내준다. 구글은 논문에서 이러한 구조를 통해 기존의 전처리 방식인 8개의 마이크를 이용한 빔포밍(Beamforming)을 적용한 후 음성인식을 하는 방법에 비해 10%정도 개선된 오인식률을 얻었다고 발표하였다.

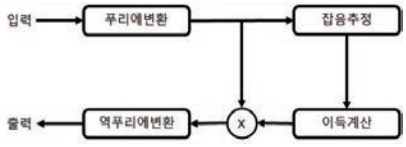
최근 구글에서는 마이크 신호를 입력으로 하는 음성인식 신경망을 발표하였는데, 빔포밍을 적용한 후 음성인식을 하는 방법에 비해 10%정도 개선된 오인식률을 얻었다고 발표하였다.

2. 음성향상

음성향상은 열화된 음성신호를 신호처리 기법을 이용하여 음질이나 음성의 명료도를 높이는 것을 의미한다. 그중에 가장 많이 연구되는 것이 잡음제거이므로 일반적으로 음성향상이라고 하면 잡음제거를 의미하는 경우도 많이 있다. 신호처리에서 가장 오래된 연구분야이지만 완전한 해법이 존재하지 않아 여전히 많이 연구되고 있다.

전통적인 방식의 잡음제거로는 스펙트럼 감산(spectral subtraction) 방법과 위너 필터(Wiener filter) 방법이 있다. 스펙트럼 감산 방법은 입력 스펙트럼에서 잡음을 추정하고 추정된 잡음의 스펙트럼을 빼서 잡음이 제거된 스펙트럼을 얻은후 이를 다시 시간영역 신호로 복원하는 것이다. 음성신호는 음성이 없는 묵음구간이 많이 존재하므로 잡음이 정상(stationary) 신호라고 가정하고 이 구간에서 잡음의 스펙트럼을 추정한다. 위너 필터 방법은 예측된 잡음의 스펙트럼과 이를 통해 예측된 음성신호의 스펙트럼을 이용하여 위너 이득(Wiener gain)을 계산하여 입력 스펙트럼에 이 이득을 곱해서 잡음제거된 스펙트럼을 얻는다. 수식(1)은 위너이득을 계산하는 식이고, 수식(2)는 스펙트럼 감산 방법의 이득을 계산하는 방법이다.

$$Gw = \frac{\text{음성파워스펙트럼}}{\text{음성파워스펙트럼} + \text{잡음파워스펙트럼}} \quad (1)$$



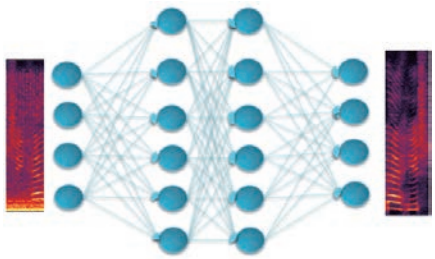
〈그림 3〉 전통적인 잡음제거 방법

$$Gs = 1 - \frac{\text{잡음스펙트럼}}{\text{입력스펙트럼}} \quad (2)$$

두 방법 모두 〈그림 3〉과 같은 과정으로 표현할수 있다. 이후 많은 연구들은 이와 비슷한 맥락에서 잡음을 정확하게 추정하는 방법이나 이를 이용하여 음질을 최대한 높힐 수 있도록 이득을 계산하는 방법들이라고 볼 수 있다.

그러나 이러한 방법들은 잡음이 정상 신호라고 가정하므로 비정상(non-stationary) 잡음에 대해서는 성능이 떨어질수 밖에 없는 한계가 있었다. 이를 극복하기 위한 방법으로 먼저 여러개의 마이크를 사용하여 빔포밍과 같은 공간 필터링을 하는 방법이 시도되었다. 이는 특정 방향의 소리를 증폭하고 다른 방향의 소리는 감쇄시키는 필터로 특정방향의 음성 신호만을 추출하기 위한 방법이다. 그리고 1990년대 말에서 2000년대 초에는 입력된 여러개의 마이크 신호에서 스스로 학습하여 원본 신호를 분리해내는 암묵 신호 분리(Blind Source Separation)에 대한 연구도 활발히 이루어졌다[10]. 이 때 비교사(unsupervised) 기계학습의 일종인 독립 요소 분석(Independent Component Analysis)[11]을 이용한 방법이 널리 적용되었다. 그러나 이 방법은 특정 조건에서 아주 잘 동작하였지만 마이크의 갯수보다 음원의 숫자가 많을 때나 잡음이 특정위치에서 나지 않고 분산되어 있을 경우 어려움이 있었다. 그리고 분리된 음원이 많을 경우 학습을 통해 분리된 음원중 원하는 음원을 특정하기 힘든 문제, 실시간 시스템에서 학습이 빠르고 안정적으로 수렴하기 힘든 문제 등으로 인해 서 현재는 제한적인 용도로만 사용되고 있는 실정이다.

전통적인 잡음제거 방법은 음질은 어느정도 향상시키지만 음성의 명료도는 거의 개선이 없다는 점이다. 이에 반해 기계학습에 의한 방법들은 음성의 명료도를 향상시킨다는 것을 알 수 있다.



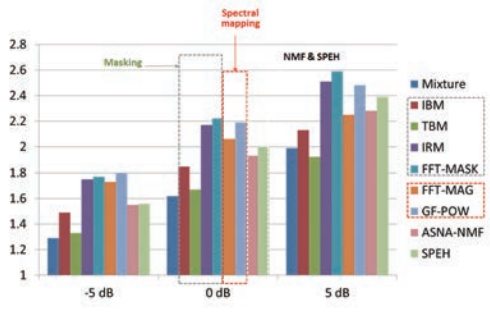
〈그림 4〉 심층 신경망을 이용한 잡음제거

이 후 하나의 마이크를 사용해서 비정상 잡음을 제거하거나 음원을 분리하고자 하는 연구가 좀 더 활발히 진행되었다. 이를 위한 방법으로는 NMF(Nonnegative Matrix Factorization)이 근래까지도 널리 사용되고 있다. NMF는 실제 신경망을 구성하는 뉴런의 발화율(firing rate)이 양수로만 되어있다는 점에서 영감을 얻어 입력 신호를 기저벡터(basis vectors)와 그에 해당하는 계수(coefficients)들로 분리해 표현하고, 이들을 모두 양수로만 구성하는 방법이다. 이를 통해 D. Lee와 H Seung은 사람의 얼굴 사진에서 눈, 코, 입과 같은 객체를 사전 정보 없이 알고리즘이 스스로 분리해 내는 결과를 얻었다[12]. 이 NMF를 이용하여 음성과 잡음에 대한 기저백터를 데이터베이스에서 미리 학습하여 입력 스펙트럼을 음성과 잡음으로 분리해 낸다[13].

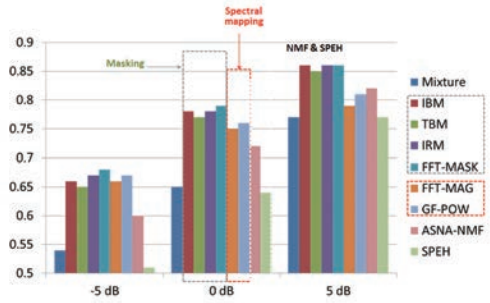
하나의 마이크를 통해서 잡음을 제거하는 또 다른 방법으로 신경망을 이용하는 방법이 있다. 〈그림 4〉와 같이 신경망의 입력으로 잡음이 섞인 음성이 들어가고 출력으로 잡음이 제거된 음성이 나오도록 신경망을 학습시키는 방법이다. 이를 잡음제거 오토인코더(denoising auto-encoder)라고 부른다[14]. 이 방법을 통해서 전통적인 잡음제거 방법에 비해 월등히 높은 PESQ(Perceptual Evaluation of Speech Quality) 점수를 얻을수 있었다.

신경망을 이용해서 잡음을 제거하는 다른 방법으로는 CASA(Computational Auditory Scene Analysis)의 접근법이 있다. 이는 잡음 제거된 음성의 스펙트럼을 직접 추정하기보다는 스펙트럼 상에서 음성에 해당하는 부분과 잡음에 해당하는 부분을 구분하는 마스커(masker)를 추정하는 방법이다[15]. 이를 추정하기 위하여 예전에는 음성의 피치 정보를 찾거나 하는 다양한 방법이 시도되었으나 최근에는 신경망을 통해서 마스커를 학습하는 방식이 주로 연구되고 있다[16].

〈그림 5〉와 〈그림 6〉은 각각 잡음제거 방법별 PESQ 점수와 STOI(short-time objective intelligibility)점수

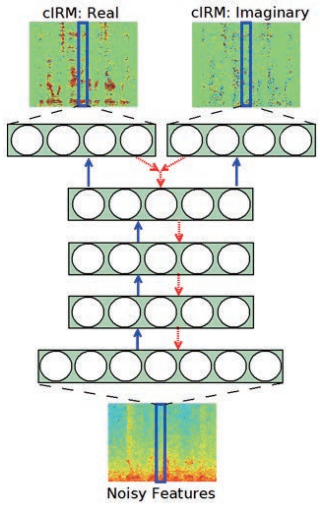


〈그림 5〉 잡음제거 방법별 PESQ점수 비교



〈그림 6〉 잡음제거 방법별 STOI점수 비교

를 비교한 결과이다. PESQ점수는 음질을 나타내는 척도이고, STOI점수는 음성의 명료도를 나타내는 척도이다. Mixture로 표시된 것이 잡음이 섞인 입력 신호의 점수이고, SPEH로 표시된 방법이 전통적인 잡음제거 방법중 가장 최근방법의 결과이다. ASNA-NMF라고 표시된 방법은 NMF를 이용한 방법중 가장 최근의 결과이고 FFT-MAG, GF-POW라고 표시된 방법은 스펙트럼을 직접 추정하는 방법이다. 그리고 나머지는 각각 다른 종류의 마스커를 추정하는 방법을 표시한다. 여기서 주목할점은 첫째, 전통적인 잡음제거 방법은 음질은 어느정도 향상시키지만 음성의 명료도는 거의 개선이 없다는 점이다. 이에 반해 기계학습에 의한 방법들은 음성의 명료도를 향상시킨다는 것을 알 수 있다. 둘째, NMF를 이용한 방법보다 신경망을 이용한 방법의 성능이 더 좋다는 점이다. 그리고 마지막으로 마스커의 종류에 따라 차이는 있지만, 스펙트럼을 직접 추정하는 방법보다 마스커를 추정하는 방법의 성능이 더 낫다는 점이다.



〈그림 7〉 신경망을 통한 복소수 마스터 예측 모델

신경망 모델은 데이터와 학습할 타겟이 주어지면 뭐든 일단 학습해볼수 있다는 장점이 있다. 그래서 최근에는 기존의 잡음제거에서 관심을 많이 가지지 않았던 원본 신호의 위상 정보를 복원하고자 하는 연구도 진행되고 있다. D. Williamson은 〈그림 7〉과 같이 원본 신호의 스펙트럼을 실수부와 허수부로 나누어서 각각을 복원 할수 있는 마스커를 학습시켰다. 그 결과 스펙트럼의 크기만을 복원한 것에 비해 비슷한 명료도는 유지하면서 PESQ점수는 10%정도 향상된 결과를 얻어 보다 좋은 음질의 잡음제거 결과를 얻었다고 보고하였다[17].

Ⅲ. 결 론

본 논고에서는 기계학습이 음성 신호처리에 쓰이고 있는 연구동향을 살펴보았다. 기계학습 분야에서 최근 가장 각광 받고 있는 딥러닝을 이용한 많은 연구들이 음성인식과 음성향상에 이용되어 기존에 오랜 기간 동안 이루지 못한 성능향상을 이루어낸 사례들을 살펴보았다. 사실 지금의 딥러닝은 1980년대에 제안된 신경망과 근본적으로 달라진 것은 없다고 볼수 있다. 하지만 이렇게 전통적인 신호처리 분야의 성능향상에 돌파구를 마련할 수 있었던 것은 예전과는 비교할수 없을만큼 많은 데이터와 이를 빠른 속도로 처리할 수 있는 컴퓨터의 계산능력 향상에 기인한다고 볼 수 있다. 그리고 이와 더불어 주어진 문제에 적합한 형태의 신경망 구조와 학습 방법을 잘 찾아낸 결과라고 볼 수도 있다. 최신 연구들의 동향으로 볼때, 음성 신호처리 분야에서 기계학습이 이용되면서 연구자

들이 기존에 가지고 있던 신호처리 지식으로 직접 설계한 알고리즘들이 신경망의 구조만 주어지고 그 신경망의 동작은 데이터를 통해서 스스로 학습되어지는 알고리즘에 의해 대체되고 있다. 향후 많은 연구자들이 기존의 신호처리 지식을 기계학습 분야에 잘 접목하여 새로운 신경망 구조와 그에 필요한 효과적인 학습 방법을 개발하여 더 많은 응용분야에서 획기적인 연구성과를 보여줄 수 있을 것이라 기대된다.

참 고 문 헌

[1] 김기응, 알파고 충격에서 무엇을 배워야 할 것인가, 중앙일보 시론, 2016년 3월 15일, http://news.joins.com/article/19723142

[2] Mastering the game of Go with deep neural networks and tree search, Nature 529, 2016년

[3] G. Hinton, Y. Bengio, Y. LeCun, NIPS tutorial: Deep Learning, 2015년

[4] A. Mohamed, D. Yu, L. Deng, Investiation of full-sequence training of deep belief networks for speech recognition, Interspeech, 2010년

[5] G. Dahl, D. Yu, L. Deng, A. Acero, Large vocabulary continuous speech recognition with context-dependent DBN-HMMS, ICASSP, 2011년

[6] H. Sak, A. Senior, K Rao, A. Graves, F. Beaufays, J Schalkwyk, Learning acoustic frame labeling for speech recognition with recurrent neural networks, ICASSP 2015년

[7] D. Amodei et al, Deep Speech 2: End-to-end speech recognition in English and Mandarin, arXiv:1512.02595, 2015년

[8] G. Saon, H.-K Kuo, S. Rennie, M. Picheny, The IBM 2015 English conversational telephone speech recognition system, ICASSP 2015

[9] T. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, M. Bacchiani, Factored spatial and spectral multichannel raw waveform CLDNNS, ICASSP 2016년

[10] J. F. Cardoso, Blind signal separation: statistical principles, Proceedings of the IEEE, 1998년

[11] A. Hyvarinen, E. Oja, Independent component analysis: algorithms and applications, Neural Networks, 2000년

[12] D. Lee, H. Seung, Learning the parts of objects by non-negative matrix factorization, Nature, 1999년

[13] P. Smaragdis, Probabilistic decompositions of spectra for sound separation, Blind speech separation, 2007년

[14] X. Lu, Y. Tsao, S. Matsuda, C. Hori, Speech enhancement based on deep denoising autoencoder, Interspeech, 2013년

[15] G. Hu, D. Wang, Monaural speech segregation based on pitch tracking and amplitude modulation, IEEE Transactions on Neural Networks, 2004년

[16] Y. Wang, A. Narayanan, D. Wang, On training targets for supervised speech separation, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014년

[17] D. Williamson, Y. Wang, D. Wang, Complex ratio masking for monaural speech separation, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016년

김 태 수

- 2001년 2월 한양대학교 전자전기공학 학사
- 2003년 2월 KAIST 전기및전자공학 석사
- 2007년 2월 KAIST 바이오및뇌공학 박사
- 2004년 4월~2006년 2월 UCSD Institute for Neural Computation, 방문연구원
- 2007년 3월~2010년 3월 LG전자기술원, 선임연구원
- 2010년 3월~현재 한국퀄컴연구소, 수석연구원

〈관심분야〉
기계학습, 신경망, 멀티미디어 신호처리