한국음향학회지 제25권 제6호 pp. 277~281 (2006)

분산 시간지연 회귀신경망을 이용한 피치 악센트 자동 인식

Automatic Recognition of Pitch Accent Using Distributed Time-Delay Recursive Neural Network

김 성 석*
(Sung-Suk Kim*)

*용인대학교 컴퓨터정보학과

(접수일자: 2006년 7월 28일; 수정일자: 2006년 8월 22일; 채택일자: 2006년 8월 24일)

본 논문에서는 시간지연 회귀신경회로망을 이용한 음절 레벨에서의 피치 악센트 자동 인식 방법을 제안한다. 시간지연 회귀 신경회로망은 두 종류의 동적 문맥정보를 표현한다. 시간지연 회귀신경회로망의 시간지연 입력 노드는 시간 축상의 피치 및 에너지 궤도를 표현하고, 회귀 노드는 피치 악센트의 특성을 반영하는 문맥 정보를 표현한다. 본 논문에서는 이러한 시간지연 회귀신경회로망을 두 가지 형태로 구성하여 피치 악센트 자동 인식에 적용한다. 하나의 형태는 단일 시간지연 회귀 신경회로망에서 복수 개의 운율 특징파라미터 (피치, 에너지, 지속시간)를 입력 노드에 함께 공급하여 피치 악센트 인식을 수행하고, 다른 하나는 분산 시간지연 회귀 신경회로망을 이용하여 피치 악센트 인식을 수행한다. 분산 시간지연 회귀 신경회로망은 여러 개의 시간지연 회귀 신경회로망으로 구성되고, 각 시간지연 회귀 신경회로망은 단일 운율 특징 파라미터만으로 학습된다. 분산 시간지연 회귀 신경회로망의 인식결과는 개별 시간지연 회귀 신경회로망의 출력 값의 가중치 합으로 결정된다. 화자 독립 피치 악센트 인식 실험을 위해 보스톤 라디오 뉴스 코퍼스(BRNC)를 사용하였다. 실험결과, 분산 시간지연 회귀 신경회로망은 83.64%의 피치 악센트 인식률을 보였다.

핵심용어: 피치 악센트, 운율, 시간지연 회귀 신경회로망, 분산 시간지연 회귀 신경회로망

투고분야: 음성신호처리분야 (2.5)

This paper presents a method for the automatic recognition of pitch accents over syllables. The method that we propose is based on the time-delay recursive neural network (TDRNN), which is a neural network classifier with two different representation of dynamic context: the delayed input nodes allow the representation of an explicit trajectory F0 (t) along time, while the recursive nodes provide long-term context information that reflects the characteristics of pitch accentuation in spoken English. We apply the TDRNN to pitch accent recognition in two forms: in the normal TDRNN, all of the prosodic features (pitch, energy, duration) are used as an entire set in a single TDRNN, while in the distributed TDRNN, the network consists of several TDRNNs each taking a single prosodic feature as the input. The final output of the distributed TDRNN is weighted sum of the output of individual TDRNN. We used the Boston Radio News Corpus (BRNC) for the experiments on the speaker-independent pitch accent recognition. The experimental results show that the distributed TDRNN exhibits an average recognition accuracy of 83.64% over both pitch events and non-events.

Key words: Pitch accent, Prosody, TDRNN, Distributed TDRNN

ASK subject classification: Speech Signal Processing (2.5)

I. 서 론

대화 중에 의미적으로 중요한 단어를 상대적으로 돋들리게 하기 위하여 화자는 이 단어에 강세 (악센트)를 두어 말한다. 이러한 강세 단어에는 현저히 높거나 낮은 기본주파수 (fundamental frequency, F0)가 발생 된다 [1]. 이런 현저히 높거나 낮은 F0 콘투어 (contour)를 피

책임저자: 김 성 석 (sskim@yongin.ac.kr)
경기도 용인시 삼가동 470번지 용인대학교 컴퓨터정보학과

치 악센트 (pitch accent)라 부르며, 피치 악센트는 음성 신호의 여러 변화들을 유발시킨다. 예를 들면, 모음의 길이와 폐쇄 자음 (stop consonant)의 유성음 시작점 (voice onset time, VOT)을 증가시킨다. Cole은 악센트가 없는 /p/의 유성음 시작점은 악센트를 동반하는 /b/의 유성음 시작점과 매우 유사함을 보였다 [2]. 따라서 피치 악센트의 위치 정보는 음성인식에서 매우 유용한 사전 정보로 사용될 수 있다.

피치 악센트와 같은 운율 정보는 다음의 이유로 음성인식에 매우 유용하다. 운율 (prosody)은 문법과 연관된다. Price는 운율 정보가 동일 음소 열로 구성된 다른 문장들을 문법적으로 애매하지 않게 하는데 사용될 수 있음을 보였고 [3], Kim은 운율 정보가 인식된 문장의 구두점을 추측하는데 사용될 수 있음을 보였다 [4]. 운율은 의미와도 연관된다. Taylor는 dialog act labels 인식을 위해 운율 정보를 사용하였고 [5], 말더듬 (speech disfluency)의 검출과 처리를 위해서도 운율 정보를 사용하였다 [6]. 또한 운율은 음소를 올바르게 레이블 하기 위한 사전 조건 정보로도 사용된다. 따라서 피치 악센트 인식은 올바른 음성인식을 위해 필요하다.

피치 악센트 인식에 관한 연구는 주로 은닉마르코프모델 (HMM)을 이용하고 있다. Taylor는 피치 악센트 인식을 위하여 혼합 가우시안 HMM 모델을 사용하였고, 보스톤 라디오 뉴스 코퍼스 (Boston Radio News Corpus, BRNC)를 이용한 화자 독립 피치 악센트 인식실험에서 72.2%의 인식률을 얻었다 [7]. 또한, Ostendorf는 HMM 기반 악센트 인식 실험에서 BRNC를 이용하여 89%의 화자 종속 피치 악센트 인식률을 얻었다 [8].

피치 악센트의 인지 특성은 비선형성과 문맥 의존성을 가지고 있으며, 이는 회귀 신경회로망으로 학습과 처리가 가능하므로 본 논문에서는 HMM에 기반 하지 않는 비모수 (non-parametric), 문맥 의존 (context-dependent) 신경회로망 모델을 사용한다. 사용된 모델은 시간지연 회귀 신경회로망 (time-delay recursive neural network, TDRNN) [9]과 본 논문에서 제안하는 분산 시간지연 회귀 신경회로망 (distributed TDRNN)이다. 피치 악센트 인식을 위해 피치 (fundamental frequency, F0), 에너지 (energy), 지속시간 (duration) 데이터를 운율 특징 파리미터로 사용하였다. 본 논문에서는 복수 개의 운율 특징 파라미터 (F0, 에너지, 지속시간)를 함께 입력으로 사용하는 TDRNN과 단일 운율 특징 파라미터만으로 학습된 여러 개의 TDRNN의 출력 결과들을 가중치 조합하여 최종 인식하는 분산 TDRNN을 구성하고, 그 성능을 비교, 분석한다. 본 논문은 다음과 같이 구성된다. 제 2장에서는 피치 악센트 인식을 위해 사용되는 TDRNN의 구조를 설명하고, 제 3장에서는 분산 TDRNN의 구조를 설명한다. 그리고 제 4장에서는 보스톤 라디오 뉴스 코퍼스를 이용한 화자 독립 피치 악센트 인식 실험결과를 기술하고 제 5장에서는 결론을 내린다.

II. 시간지연 회귀 신경회로망

본 절에서는 피치 악센트 인식을 위한 시간지연 회귀 신경회로망 (TDRNN)의 구조와 학습 방법에 관하여 기술한다. 그림 1은 TDRNN의 구조를 보이고 있다. TDRNN은 다층퍼셉트론 신경회로망에 피치문맥층 (pitch context layer)이 부가되었다 [9]. 이러한 TDRNN은 두 종류의 동적 문맥 (dynamic context) 정보를 표현하는 수단을 제공한다. 시간 지연된 입력층 노드들은 (time-delayed input units) 시간 축 상의 피치 (F0) 및 에너지 궤도의 단기 문맥 (short-term context) 정보를 표현하고, 피치문맥층를 통한 다중 회귀 루프들은 피치 악센트의 특성을 표현하는 F0 궤도의 장기 문맥 (long-term context) 정보를 표현한다. 문맥정보는 피치 악센트의 올바른 레이블을 위해 매우 중요하다. 왜냐하면 단어 (word)는 단지 하나의 주된 강세 음절 (accented syllable)을 가지고 있고, 이 주된 강세 음절은 주변 다른 음절에 영향을 미치기 때문이다. 학습 시간 t에서 피치층 (pitch layer) 노드의 출력은 피치문맥층 노드로 복사 (copy)되어 다중 지연되고, 시간 t+1에서 F0 (t+1)과 함께 입력으로 작용한다. 이러한 TDRNN은 통상적인 오류 역전파 (error back-propagation) 학습 알고리즘 [10]에 의해 학습된다.

그림 1에서 입력층 (input layer) 노드와 은닉층 (hidden layer) 노드 사이의 연결과 피치문맥층 노드와 은닉층 노드 사이의 연결을 표시하는 시간지연 박스 (delay box)의 구조는 그림 2와 같다. 그림 2에서, 층 $h-1$ 의 노드 i 는 상위 층 h의 노드 j에 연결되고, 각 연결 라인 (connection line)은 개별적인 시간지연 $\tau_{ijk,h-1}$와 연결강도 (weight) $w_{jik,h-1}$을 가진다. 각 노드의 출력 값은 식 (1)과 같이 하위 층 노드들의 출력 값의 가중치 합으로 표현된다.

$$x_{j,h}(t_n) = f\Big(\sum_{i \in N_{h-1}} \sum_{k=1}^{K_{ij,h-1}} w_{jik,h-1} \cdot x_{i,h-1}(t_n - \tau_{ijk,h-1})\Big) \quad (1)$$

식 (1)에서 $x_{j,h}(t_n)$는 발화 문장 가운데 t_n번째 음절에서 층 h의 노드 j의 출력 값이다. 그리고 $x_{i,h-1}(t_n-\tau_{ijk,h-1})$ 은 $t_n-\tau_{ijk,h-1}$번째 음절에서 층 $h-1$의 노드 i의 활성화 레벨이고, N_{h-1} 는 층 $h-1$에 있는 전체 노드들의 수를 표시한다. $K_{ji,h-1}$은 층 $h-1$의 노드 i로부터 층 h의 노드 j에 연결된 전체 연결 라인의 수 (즉, 시간 지연 회수)를 나타낸다. $f(.)$ 는 시그모이드 (sigmoid) 활성화 함수이다. 연결강도 $w_{jik,h-1}$은 통상의 오류 역전파 학습 알고리즘 [10]에 의해 학습되고 시간지연 $\tau_{ijk,h-1}$은 학습되지 않고 고정된 시간지연 값을 가진다. 그리고 복사 (copy) 오퍼레이션을 통한 귀한 연결 (feedback connection)은 학습되지 않고 1로 고정된다.

TDRNN는 두 가지 형태로 구성되어 피치 악센트 자동 인식에 사용된다. 하나의 형태는 단일 TDRNN에서 3 개의 운율 특징 파라미터 (피치, 에너지, 지속시간)를 입력 노드에 모두 함께 공급하여 피치 악센트 인식을 수행하는 형태이고 (normal TDRNN), 이때 입력 노드는 수는 3이다. 그리고 다른 하나는 1개의 운율 특징 파라미터만으로 학습된 3개의 TDRNN (입력 노드 수는 1)의 출력 결과들을 가중치 조합하여 인식을 수행하는 형태 (distributed TDRNN) 이다. TDRNN의 입력 노드에 공급되는 값은 음절단위 운율 특징 파라미터이다. 음절 단위 운율 특징 파라미터는 음절 내에 있는 프레임들의 특징 파라미터를 식 (2)와 같이 평균하여 구하였다.

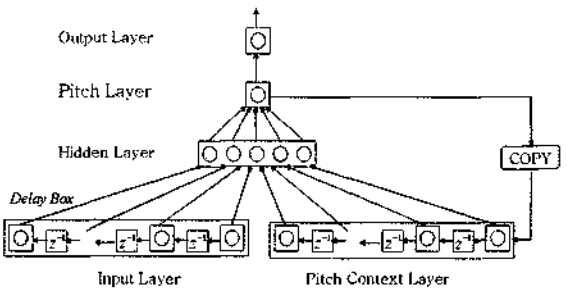

그림 1. 시간지연 회귀 신경회로망의 구조 (Z^{-1}는 한 음절 시간 지연을 나타낸다)

Fig. 1. The architecture of TDRNN (Z^{-1} denotes one syllable time delay).

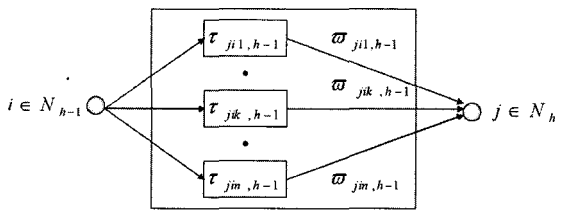

그림 2. 시간 지연 박스 ($\tau_{ijk,h-1}$은 층 $h-1$ 의 노드 i 로부터 층 h 의 노드 j 사이의 k번 음절 시간지연을 표시한다)

Fig. 2. Delay box ($\tau_{ijk,h-1}$ denotes k th syllable time delays to node j of layer h from node i of layer $h-1$).

$$D_m=\frac{1}{N_{\phi_m}}\sum_{f\in\phi_m}F_m[f] \qquad (2)$$

식 (2)에서 D_m은 음절 S_m의 특징 파라미터이고, $F_m[f]$은 음절 S_m에 속한 프레임 f의 특징 파라미터이다. 그리고 ϕ_m은 음절 S_m에 속한 프레임들의 집합이며, N_{ϕ_m}은 ϕ_m에 있는 총 프레임 수이다. 분산 시간지연 회귀 신경회로망 (distributed TDRNN)의 구조는 다음 절에서 설명된다.

III. 분산 시간지연 회귀 신경회로망

피치 악센트 인식에는 피치 (F0), 에너지, 지속시간이 운율 특징 파라미터로 널리 사용된다. 이러한 개별 운율 특징 파라미터는 피치 인식기 성능에 각각 다르게 기여함을 표1에서 알 수 있다. F0 운율 특징 파라미터만을 이용한 TDRNN (pitch-based TDRNN)의 성능이 다른 운율 특징 파라미터를 이용한 TDRNN 보다 우수함을 알 수 있고, F0가 피치 악센트 인식에서 가장 중요한 운율 특징 파라미터임을 확인 할 수 있다. 그런데 모든 운율 특징 파라미터 (피치, 에너지, 구간)를 입력 노드에 모두 함께 공급하여 피치 악센트 인식을 수행하는 TDRNN (normal TDRNN)은 피치 악센트 인식에서 운율 특징 파라미터 간의 각기 다른 기여도를 무시하는 망 구조이다. 따라서 각기 다른 운율 특징 파라미터로 학습된 독립된 TDRNN의 결과를 조합하여 인식하는 형태의 망 구조가 요구됨을 알 수 있다.

분산 시간지연 회귀 신경회로망 (distributed TDRNN)은 각기 다른 운율 특징 파라미터를 다른 TDRNN으로

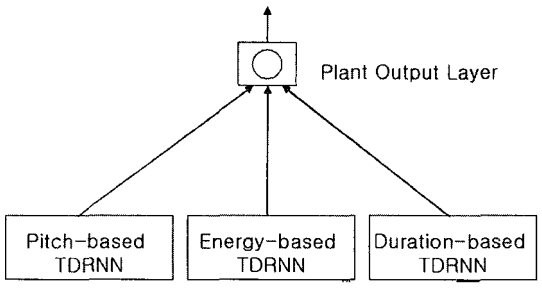

그림 3. 분산 시간지연 회귀 신경회로망의 구조

Fig. 3. The architecture of distributed TDRNN.

모델링하고 개별 TDRNN의 결과 값을 가중치 조합하여 인식하는 망 구조이다. 그림 3은 분산 TDRNN의 구조를 보이고 있다. 분산 TDRNN에서 플란트 출력층 (plant output layer)의 최종 출력 값은 개별 TDRNN의 출력 값의 가중치 합이다. 이러한 분산 TDRNN에서 개별 TDRNN은 개별 운율 특징 파라미터에 최적인 망 파라메터 (시간지연 수, 은닉 노드 수)를 가질 수 있게 설계될 수 있고, 개별 운율 특징 파라미터의 기여도는 학습에 의해 최적으로 조정되어 피치 악센트의 인식 성능을 개선시킨다.

IV. 실험 결과

화자 독립 피치 악센트 인식 실험을 위하여 Boston Radio News Corpus (BRNC)를 사용하였다. BRNC는 7명의 전문 라디오 아나운서가 라디오 뉴스 스토리를 녹음한 음성 코퍼스로 국제적으로 널리 이용되고 있다 [11]. BRNC는 ToBI (tone and break indices) [12] 운율 표기법을 이용하여 피치 악센트와 경계 (boundary)를 표기하고 있다. 본 논문에서는 high 악센트 (H*), downstepped 악센트 (!H*), low 악센트 (L*), questionable 악센트 (?*), 악센트 타입이 표기되지 않은 악센트 (*, 보통 high 또는 downstepped 악센트 임)로 표기된 모든 음절들을 피치 악센트를 동반하는 음절로 간주하여 실험에 사용하였다. Boston Radio News Corpus에 있는 F1A 여성 화자의 음절 단위 F0, 에너지, 지속시간을 추출하여 학습용 데이터로 사용하고, F2B 여성 화자로부터 추출한 F0, 에너지, 지속시간을 시험용 데이터로 사용하였다. F0는 Entropic XWAVES에 있는 "formant" 프로그램을 이용하여 추출되었다. 학습과 시험에 사용된 피치 악센트를 동반하는 음절의 개수는 8160 개이고, 피치 악센트를 동반하지 않은 음절의 수는 24,208 개이다.

표 1은 피치 악센트 인식 실험결과를 보이고 있다. 실험결과에서 단일 운율 특징 파라미터를 이용한 TDRNN들 가운데 피치 (F0)를 이용한 TDRNN (pitch-base TDRNN)이 에너지나 지속시간을 이용한 TDRNN보다 피치 악센트 인식률이 우수함을 보이고 있다. 실험에 사용된 피치 기반 TDRNN (pitch-based TDRNN)의 은닉 노드 수는 5, 입력 노드의 시간지연 수는 14, 피치문맥층 노드의 시간지연 수는 18 이고, 에너지 데이터만을 이용한 TDRNN (energy-based TDRNN)의 은닉 노드 수는 8, 입력 노드의 시간지연 수는 16, 피치문맥층 노드의 시간지연 수는 15 이다. 그리고 음절의 지속시간 데이터만을 이용한 TDRNN (duration-based TDRNN)의 은닉 노드 수는 10, 입력 노드의 시간지연 수는 12, 피치문맥층 노드의 시간지연 수는 14로 하였다. 표1의 결과로부터 피치 악센트 인식에서 F0 정보가 매우 중요함을 알 수 있다. 그리고 피치, 에너지, 지속시간을 모두 입력으로 하는 TDRNN (normal TDRNN)은 단일 운율 특징 파라미터를 이용한 TDRNN들 보다 피치 악센트 인식에서 우수한 성능을 보였다. 실험에 사용된 normal TDRNN의 은닉 노드 수는 10, 입력 노드의 시간지연 수는 14, 피치문맥층 노드의 시간지연 수는 18 이다. 그러나 단일 운율 특징 파라미터로 학습된 TDRNN의 출력 결과들을 가중치 조합하여 최종 인식하는 분산 TDRNN이 가장 우수한 인식 성능을 보이고 있다.

표 1. 단일 운율 특징 파라미터를 이용한 TDRNN, 복수 특징 파라미터를 이용한 TDRNN 및 분산 TDRNN의 피치 악센트 인식률
Table 1. The classification accuracy of pitch accent and non-accent over syllable using TDRNNs.

	Accent (%)	Non-accent (%)	Accent + Non-accent (%)
Pitch-based TDRNN	69.94	84.23	79.79
Energy-based TDRNN	75.92	72.26	73.40
Duration-based TDRNN	85.18	39.21	53.51
Normal TDRNN	68.27	87.05	81.21
Distributed TDRNN	78.20	86.09	83.64

V. 결 론

피치 악센트의 인식은 음성인식에서 매우 유용한 사전 정보로 사용된다. 피치 악센트 인식에 관한 대부분의 연구는 은닉마르코프모델 (HMM)을 이용하고 있다. 본 논문에서는 HMM에 기반 하지 않은 비모수 (non-parametric), 문맥의존 (context-dependent) 신경회로망 모델을 사용하여 피치 악센트 인식을 수행하였다. 사용된 신경회로망 모델은 시간지연 회귀 신경회로망 (TDRNN)이다. 그리고 화자 독립 피치 악센트 인식 실험을 위해 보스톤 라디오 뉴스 코퍼스 (BRNC)를 사용하였고, 운율 특징 파라미터로 피치 (F0), 에너지, 지속시간을 사용하였다.

본 논문에서는 TDRNN을 두 가지 형태로 구성하여 피치 악센트 인식에 적용하였다. 하나의 형태는 단일 TDRNN에서 복수 개의 운율 특징 파라미터 (피치, 에너지, 지속시간)를 입력 노드에 함께 공급하여 피치 악센트 인식을 수행하고, 다른 하나는 단일 운율 특징 파라미터로만으로 학습된 여러 개의 TDRNN의 출력 값을 가중치 조합하여 최종 인식하는 분산 TDRNN을 구성하여 피치 악센트 인식을 수행하였다. 실험결과, 개별 운율 특징 파라미터에 최적인 망 파라메터를 가지는 분산 TDRNN이 가장 우수한 피치 악센트 인식 결과를 보였다. 그리고 피치, 에너지, 지속시간의 개별 운율 특징 파라미터를 이용한 피치 악센트 인식 실험에서는 표1에서 보듯이 피치 (F0)가 에너지나 지속시간보다 피치 악센트 인식에 더욱 유용함을 알 수 있었다. 분산 TDRNN은 피치 악센트에 대하여 78.20%의 인식률을 보이고, non-accent에 대해서는 86.09%의 인식률을 보였다. 그리고 피치 악센트와 non-accent 전체에 대하여 83.6%의 인식률을 보였다. 이 결과는 Taylor가 BRNC와 혼합 가우시안 HMM 모델을 이용하여 화자 독립 피치 악센트 인식실험을 수행한 결과 (전체 인식률 72.2%)보다 우수하다 [7].

참 고 문 헌

1. Mary E. Beckman and Janet Pierrehumbert, "Intonational structure in Japanese and English," Phonology Yearbook, 3 255-309, 1986.
2. Jennifer Cole, Hansook Choi, Heejin Kim, and Mark Hasegawa-Johnson, "The effect of accent on the acoustic cues to stop voicing in radio news speech," in Proc. Inter. Conf. Phonetic 2003.
3. P.J. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong, "The use of prosody in syntactic disambiguation," J. Acoust. Soc. Am, **90** (6) 2956-2970, 1991.
4. Ji-Hwan Kim and Philip C. Woodland, "The use of prosody in a combined system for punctuation generation and speech recognition," in Proc. EUROSPEECH, 2001.
5. P. Taylor, S. King, S. Isard, H. Wright and J. Kowtko, "Using intonation to constrain language models in speech recognition," in Proc. EUROSPEECH, 1997.
6. Christine H. Nakatani and Julia Hirschberg, "A corpus-based study of repair cues in spontaneous speech," J. Acoust. Soc. Am. 95 (3) 1603-1616, 1994.
7. Paul Taylor,"Analysis and synthesis of intonation using the Tilt model," J. Acoust. Soc. Am, **107** (3) 1697-1714, 2000.
8. M. Ostendorf and K. Ross, "A multi-level model for recognition of intonation labels," in Computing prosody: computational models for processing spontaneous speech. Springer-Verlag New York, Inc., 1997.
9. Sung-Suk Kim, "Time-delay recurrent neural network for temporal correlations and prediction," Neuorcomputing, **20** 253-263, Elsevier 1998.
10. Rumelhart D. E., McClelland J. L., and the PDP Research Group, "Learning representations by back-propagating errors," in Parallel Distributed Processing, **1** 318-362. MIT Press, 1986.
11. M. Ostendorf, P.J. Price, and S. Shattuck-Hufnagel, "The Boston University Radio News Corpus," Linguistic Data Consortium, 1995.
12. Joseph F. Pitrelli, Mary Beckman, and Julia Hirschberg, "Evaluation of prosodic transcription labeling reliability in the TOBI framework," in Proc. ICSLP, 1994.

저자 약력

•김 성 석 (Sung-Suk Kim)

1985년 2월: 영남대학교 전기공학과 졸업 (공학사)
1987년 2월: 울산대학교 대학원 전자공학과 졸업 (공학석사)
1990년 8월: 울산대학교 대학원 전자 및 컴퓨터공학과 (공학박사)
1985년 3월~1991년 2월: KEPCO
2002년 미국: Carnegie Mellon University, Language Technology Institute 초청 연구원
2003년 미국: University of Illinois (Urbana-Champaign), Beckman Institute 초청 교수
1995년 3월~현재: 용인대학교 컴퓨터정보학과 부교수
※주 관심 분야: 음성인식, 신경회로망, 컴퓨터원용학습 (CALL), 음원분리