ISSN : 1226-7244 (Print)
ISSN : 2288-243X (Online)
j.inst.Korean.electr.electron.eng.Vol.24,No.4,998~1003,December 2020
논문번호 20-04-11
http://dx.doi.org/10.7471/ikeee.2020.24.4.998
78

감정 적응을 이용한 감정 인식 학습 방법

A Training Method for Emotion Recognition using Emotional Adaptation

김 원 구*★

Weon-Goo Kim*★

Abstract

In this paper, an emotion training method using emotional adaptation is proposed to improve the performance of the existing emotion recognition system. For emotion adaptation, an emotion speech model was created from a speech model without emotion using a small number of training emotion voices and emotion adaptation methods. This method showed superior performance even when using a smaller number of emotional voices than the existing method. Since it is not easy to obtain enough emotional voices for training, it is very practical to use a small number of emotional voices in real situations. In the experimental results using a Korean database containing four emotions, the proposed method using emotional adaptation showed better performance than the existing method.

요 약

본 논문에서는 기존 감정 인식 시스템의 성능 향상을 위하여 감정 적응을 사용한 감정 학습 방법이 제안되었다. 감정 적응을 위하여 적은 개수의 학습 감정 음성과 감정 적응 방식을 사용하여 감정이 없는 음성 모델로부터 감정 음성 모델이 생성되었다. 이러한 방법은 기존 방법보다 적은 개수의 감정 음성을 사용하여도 우수한 성능을 나타내었다. 학습을 위하여 충분한 감정 음성을 얻는 것은 쉽지 않기 때문에 적은 개수의 감정 음성을 사용하는 것은 실제 상황에서 매우 실용적이다. 4가지 감정이 포함된 한국어 데이터베이스를 사용한 실험 결과에서 감정 적응을 이용한 제안된 방법이 기존 방법보다 우수한 성능을 나타내었다.

Key words : emotion recognition, emotional speech, emotional adaptation, GMM, speech parameter

Ⅰ. 서론

감정은 인간 지능의 중요한 상징으로 대인 관계에서의 필수적인 요소이다. 인간은 인간의 언어를 이해하고 사람의 감정을 판단하여 자연스러운 인간-컴퓨터 상호 작용을 구현하기 위해 노력해 왔다. 현재 음성 인식 기술과 화자 인식 기술은 실용화가 가능할 정도로 안정화되어 있어 현재 미래기술로 많은 관심을 받고 있는 분야는 감정 인식 분야이다. 특히, 기계가 인간 감정을 파악하고 그에 따라 정서적으로 반응하는 기계가 개발이 개발된다면 보다 고차원적인 인간-컴퓨터 인터페이스가 가능한 제품 개발과 사용자 중심의 맞춤형 서비스가 가능할 것이다.

* Professor, Dept. of Electrical Engineering, Kunsan National University
* Corresponding author
E-mail : wgkim@kunsan.ac.kr, Tel : +82-63-469-4745
Manuscript received Nov. 26, 2020; revised Dec. 14, 2020; accepted Dec. 15, 2020.

인간은 얼굴, 몸동작, 음성이나 심장 박동수, 체온, 혈압 등의 생체신호 등의 다양한 방법으로 감정을 나타내며, 응용 분야에 따라 감정 정보를 취득하는 방법이 달라진다. 특히, 전화와 같이 음성만을 사용하는 응용 분야나 센서가 신체에 직접 접촉하지 못하는 경우에는 음성을 이용한 감정 인식 방법이 많은 이점을 가진다.

현재까지 감정 인식 분야에서 다양한 연구가 진행되어 왔다[1]. Fukuda는 음성의 템포와 에너지 정보를 사용하여 여섯 가지 감정에 대한 인식 실험을 수행하였으며[2], Moriyama는 음성의 피치와 에너지 포락선을 사용하여 감정 분류 실험을 수행하였다[3]. 또한 Silva는 얼굴 표정과 음성을 동시에 사용하는 바이모달 감정 인식 시스템을 구현하여 감정 인식 실험을 하였다[4]. Amol T.는 멜 켑스트럼 계수(MFCC)와 4가지 음성 특성을 결합한 감정 인식 실험을 수행하였다[5]. 또한 음성을 사용하여 감정 인식을 수행하기 위하여 음성의 운율, 스펙트럼, 특징 파라메터 선택 방법, 인식기 등 다양한 방법이 제안되었다[6].

국내에서도 음성 및 얼굴 표정을 사용하여 감정을 인식하는 연구가 활발하게 진행되고 있다. 우리나라 전통 국악인 창에서 인간의 희로애락을 나타내는 음의 높낮이와 장단의 특성을 분석하는 연구가 수행되었고[7], 인간의 대화 내용에서 단어, 톤, 말의 빠르기나 음질 등을 분석하여 화난 감정의 특성을 파악하는 연구도 실행되었다[8].

본 논문에서는 적은 수의 감정 음성을 사용하여 음성 감정 모델에 감정 적응을 이용한 학습 방법을 제안하였다. 기존 방법은 많은 개수의 감정 학습 데이터가 필요하였으나 본 연구에서는 감정 적응을 위하여 적은 개수의 학습 감정 음성과 화자 적응 방식을 사용하여 감정이 없는 음성 모델로부터 감정 음성 모델이 생성되었다. 우수한 성능의 감정 인식 시스템을 만들기 위하여 학습에 많은 감정 음성을 사용하는 것이 바람직하지만 학습을 위하여 충분한 감정 음성을 수집하는 것은 어려운 일이다. 따라서 적은 개수의 감정 음성을 사용하여 우수한 성능의 감정 인식 시스템을 만드는 방법은 실제 상황에서 매우 실용적이다.

본 논문의 구성은 다음과 같다. 2장에서는 감정 적응을 이용한 감정 인식 시스템의 학습 방법에 관하여 설명하고 3장에서는 감정 데이터를 사용한 감정 인식 실험을 수행하고 기존 시스템과 성능을 비교하였고 4장에서 결론을 맺는다.

II. 감정 인식 시스템의 감정 적응

본 연구에서는 감정 음성을 사용한 감성 인식 인식 시스템의 성능 향상을 위하여 감정 적응을 사용한 학습 방법이 제안되었다. 학습 과정에서 적은 개수의 학습 감정 음성과 화자 적응 방식을 사용하여 감정이 없는 음성 모델로부터 감정 음성 모델이 생성되었다. 이러한 구조를 갖는 감정 인식 시스템의 학습 과정은 그림 1과 같다.

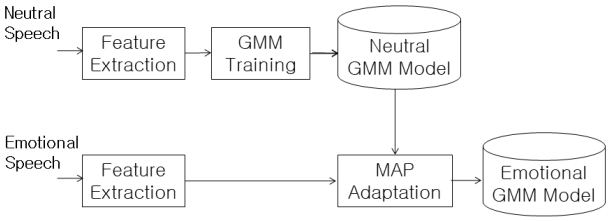

Fig. 1. Training procedure of proposed emotion recognition system.

그림 1. 제안된 감정 인식 시스템의 학습 과정

그림 1에서 각각의 감정 음성 모델을 생성하기 위하여 평상 음성(neutral speech)과 감정 음성(emotional speech)이 각각 사용된다. 여기서 감정 음성 모델로는 GMMM이 사용되었다. 우선 각 화자의 학습 데이터중에서 감정이 없는 평상 음성을 사용하여 특징 벡터를 구한다. 특징 벡터는 평상 GMM 모델의 학습 과정에 사용되고, 이렇게 생성된 평상 감정 모델은 저장된다. 그 후 각각의 감정 음성을 사용하여 특징 벡터를 구한 후 평상 감정의 GMM 모델에 감정 적응(MAP adaptation)을 수행하여 감정 GMM 모델(emotional GMM model)을 생성한다.

기존 방법들은 감정 모델을 생성하기 위하여 감정 음성만을 사용하여 모델을 만들거나 비슷한 특성을 보이는 평상 음성과 감정 음성들을 분류하고 학습 과정을 통하여 감정 모델을 생성하였다. 본 논문에서 제안된 방법은 평상 음성을 사용한 감정 모델에 적은 개수의 감정 음성을 사용하고 화자 적응을 통하여 새로운 감정 모델을 생성하는 것이다.

감정 모델을 생성하기 위한 감정 적응을 위하여 화자 적응에 널리 사용되는 베이지안 적응(Bayesian

adaptation) 또는 사후 최대 값(maximum a posteriori : MAP) 추정 방법을 사용하였다[9, 10]. 이러한 적응을 통해 평상 모델은 학습 감정 음성을 사용하여 감정 모델로 변환된다.

감정 적응을 위한 세부 과정은 다음과 같다. 먼저 평상 음성을 사용하여 GMM 기반의 평상 GMM 모델 λ를 생성한다. 이때 M개의 가우시안 분포를 사용한 혼합 분포는 다음과 같이 정의된다.

$$p(\boldsymbol{x} \mid \lambda) = \sum_{i=1}^{M} w_i p_i(\boldsymbol{x}) \tag{1}$$

가우시안 혼합 분포는 평균 벡터 μ_i, 공분산 행렬 Σ_i과 가중 w_i의 세 가지 파라메터가 사용된다. 이러한 파라메터로 GMM 모델을 나타내면 다음과 같다.

$$\lambda = \{w_i,\ \mu_i,\ \Sigma_i\},\ i = 1,...,M \tag{2}$$

두 번째로 감정 GMM 모델 $\hat{\lambda}$를 생성하기 위하여 감정 학습 음성을 사용하고 MAP 추정 방법을 적용하여 평상 GMM 모델 λ의 평균, 공분산 행렬, 가중을 변환한다. 이러한 과정은 다음과 같다.

우선 평상 GMM 모델 λ를 사용하여 감정 적응을 위한 감정 학습 음성의 특징 벡터 $X = \{\boldsymbol{x_1}, ..., \boldsymbol{x_T}\}$에 대한 확률 $P(i|\boldsymbol{x}_t)$와 통계 값들을 구한다.

$$P(i \mid \boldsymbol{x}_t) = \frac{w_i\, p_i(\boldsymbol{x}_t)}{\sum_{j=1}^{M} w_j\, p_j(\boldsymbol{x}_t)} \tag{3}$$

$$n_i = \sum_{t=1}^{T} P(i \mid \boldsymbol{x}_t) \tag{4}$$

$$E_i(\boldsymbol{x}) = \frac{1}{n_i} \sum_{t=1}^{T} P(i \mid \boldsymbol{x}_t)\boldsymbol{x}_t \tag{5}$$

$$E_i(\boldsymbol{x}^2) = \frac{1}{n_i} \sum_{t=1}^{T} P(i \mid \boldsymbol{x}_t)\boldsymbol{x}_t^2 \tag{6}$$

위의 확률과 통계 값들을 사용하여 평상 GMM 모델 λ의 i번째 혼합 분포 파라메터들을 다음과 같이 적응시켜 감정 GMM 모델 $\hat{\lambda} = \{\hat{w}_i,\ \hat{\mu}_i,\ \hat{\Sigma}_i\}$를 생성한다.

$$\hat{w}_i = [\alpha_i n_i / T + (1-\alpha_i) w_i]\gamma \tag{7}$$

$$\hat{\boldsymbol{\mu}}_i = \alpha_i E_i(\boldsymbol{x}) + (1-\alpha_i)\boldsymbol{\mu}_i \tag{8}$$

$$\hat{\boldsymbol{\sigma}}_i^2 = \alpha_i E_i(\boldsymbol{x}^2) + (1-\alpha_i)(\boldsymbol{\sigma}_i^2 + \boldsymbol{\mu}_i^2) - \hat{\boldsymbol{\mu}}_i^2 \tag{9}$$

여기서 스케일 요소 γ는 혼합 가중의 합을 1로 만들기 위한 것이고 적응 계수 α_i는 다음과 같이 정의되며 r은 실험적으로 정하는 고정된 상수이다.

$$\alpha_i = \frac{n_i}{n_i + r} \tag{10}$$

이렇게 감정 적응된 음성 인식 모델들은 감정 개수만큼 생성된다.

인식 단계에서는 입력 음성에 대한 유사도를 화자마다 S개의 평상과 감정 모델들에 대하여 식 (1)을 사용하여 다음과 같이 구한다.

$$\log p(X \mid \lambda) = \max_{1 \le k \le S} \sum_{t=1}^{T} \log p(\boldsymbol{x}_t \mid \lambda_k) \tag{11}$$

III. 실험 및 결과

3.1. 데이터베이스

감정 인식 시스템의 성능 평가를 위하여 기쁨, 슬픔, 화남과 평상을 포함한 4가지 감정을 포함한 데이터베이스를 사용하였다. 녹음은 평소 감정을 표현하는 훈련이 된 아마추어 연극단원 남/녀 각 15명을 대상으로 조용한 사무실 환경에서 이루어졌다. 각 화자는 45개의 서로 다른 문장을 4가지 감정으로 녹음하였다. 감정이 적절히 반영된 데이터베이스를 구축하기 위하여 주관적 평가를 수행하였다. 주관적 평가는 전체 5400문장을 음성 신호처리에 숙련된 연구원 10명이 청취한 후 감정이 적절히 반영되었다고 판단되는 문장을 선택하였다. 이런 과정을 통하여 총 5400개의 음성 중에서 2237개의 음성을 선별하여 최종 데이터베이스로 구성하였다.

3.2. 특징 파라메터

화자 인식을 위하여 감정이 포함된 음성 신호로부터 MFCC 파라메터를 추출하여 사용하였다. MFCC 파라메터는 다음과 같은 과정을 통하여 추출되었다. 음성 신호는 전처리 과정을 통하여 16kHz, 16비트로 샘플링되고, 고주파 성분이 보강되었다. 이렇게 샘플링된 신호는 에너지 파라메터를 사용하는 음성구간 검출 과정을 통해 묵음 구간이 제거되었다. 검출된 음성 신호로부터 20ms (320샘플)의 길이를 갖는 해밍 창을 사용하고 10ms씩 이동하면서 12차의 MFCC 파라메터가 생성되었다. 특징 파라메터

의 시간적인 변화에 대한 정보를 포함하는 델타 켑스트럼도 사용하였다.

3.3. 감정 인식 시스템 구성

본 연구에서는 감정 적응을 이용한 감정 인식 학습 방법의 성능 평가를 위하여 GMM 기반의 화자 및 문장 독립 감정 인식 시스템을 구현하였다(그림 1).

그림 1에서 GMM 모델의 학습을 위하여 20명(남성 10명, 여성 10명)이 총 45개의 문장 중에서 35개의 문장을 녹음한 음성이 사용되었다. 인식에는 학습에 참여하지 않은 10명(남성 5명, 여성 5명)을 학습에 사용되지 않은 나머지 10개의 문장을 녹음한 음성이 사용되었다.

제안된 시스템과의 성능 비교를 위하여 4가지 감정 데이터를 사용하여 감정별 모델을 생성하여 감정 인식을 수행하는 기존 시스템을 구현하였다[11]. 감정 학습 과정에서 감정 음성은 감정마다 500 문장 이상 사용되었다. 제안된 감정 적응을 위해서는 평상 감정의 GMM을 이용하여 감정마다 20~300 개까지 변경하면서 감정 GMM 모델을 생성하였다. 본 연구에서 제안된 방법과 성능 비교를 위하여 다음과 같이 시스템을 구현하였다.

1) 기존 시스템 : 학습 과정에서 감정별 음성을 사용하여 개별 감정별 모델을 생성[11]

2) 평상 모델로부터 감정 적응(제안된 방법) : 평상 음성을 사용한 GMM에 감정 음성을 사용하여 감정 적응하여 모델 생성

3.4. 실험 결과

본 실험에서는 기존 감정 인식 시스템에 사용된 켑스트럼 파라메터를 사용한 감정 인식 시스템의 성능을 평가하였다. 그림 2는 MFCC를 각각 사용하여 4가지 감정에 대하여 감정 인식을 수행한 결과를 나타낸다. 켑스트럼 파라메터로는 MFCC, ΔMFCC, ΔΔMFCC를 연결하여 사용하였다. 그림에서 알 수 있듯이 MFCC에 ΔMFCC와 ΔΔMFCC를 결합하여 사용할수록 인식 성능이 상승되어 MFCC+ΔMFCC+ΔΔMFCC의 경우에 가장 우수한 71.6%의 인식 성능을 나타내었다.

표 1은 켑스트럼 파라메터로 MFCC+ΔMFCC+ΔΔMFCC를 사용한 경우의 감정 인식 시스템의 감정별 인식률을 나타내는 표이다. 표에서 '평상' 감정의 인식은 73.7%, '기쁨' 감정의 인식은 66.0%,

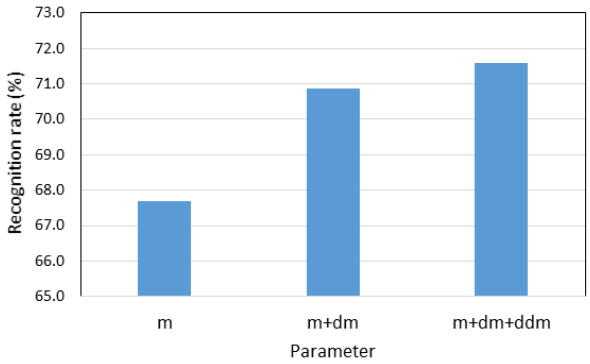

Fig. 2. Emotion recognition results using MFCC.
그림 2. MFCC 파라메터를 사용한 감정 인식 결과 (m : MFCC, m+dm : MFCC+ΔMFCC, m+dm+ddm : MFCC+ΔMFCC+ΔΔMFCC)

'슬픔' 감정의 인식은 74.7%, '화남' 감정의 인식은 71.9%로서 최종 인식률은 71.6%이다. '평상'과 '슬픔' 감정 인식은 우수한 편이나, '기쁨' 감정은 '화남' 감정과의 구분이 명확하지 않았다.

Table 1. Recognition performance of emotion recognition system.
표 1. 감정 인식 시스템의 인식 성능

	recognition rate(%)			
emotion	neutral	happy	sad	angry
neutral	73.7	15.8	3.9	6.6
happy	4.0	66.0	8.0	22.0
sad	14.3	11.0	74.7	0.0
angry	3.5	22.8	1.8	719
average	71.6			

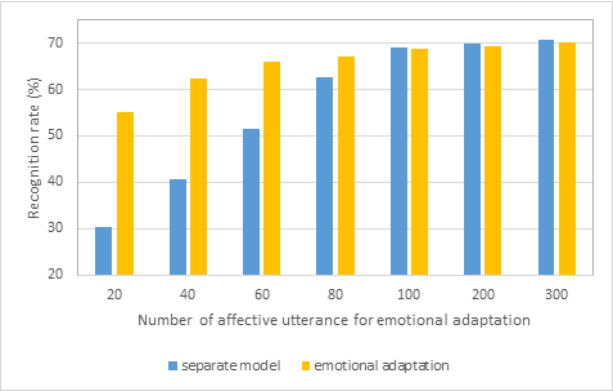

Fig. 3. Performance comparison of emotion recognition systems according to the number of affective utterance per emotion.
그림 3. 학습에 사용된 감정별 감정 발성 개수와 학습 방법에 따른 감정 인식 시스템의 성능 비교

다음 실험에서는 평상 음성을 사용한 GMM에 감정 음성을 사용하여 감정 적응하여 감정 모델 생성한 제안된 학습 방법에 대한 감정 인식 실험을 수행하였다. 그림 3은 학습 과정에서 사용된 감정 음성의 개수에 다른 감정 인식 시스템의 성능을 비교하였다. 그림에서 기존 시스템(seperate model)은 학습 과정에서 감정별 음성을 사용하여 개별 감정별 모델을 생성하였기 때문에 학습에 사용된 감정 음성의 개수가 적을수록 성능 저하가 크게 발생한다. 그러나 제안된 방법(emotional adaptation)은 평상 음성을 사용한 GMM에 감정 음성을 사용하여 감정 적응하여 모델을 생성하였기 때문에 감정 적응에 사용된 감정 음성의 개수가 적을 때도 기존 시스템에 비하여 우수한 성능을 나타내었다. 감정 적응에 사용된 감정 음성의 개수가 감정마다 100개 정도 사용되어도 기존 시스템의 성능에 거의 근접한 것을 볼 수 있다. 실제 상황에서는 화자로부터 학습을 위하여 충분한 감정 음성을 얻는 것이 쉽지 않기 때문에 적은 개수의 감정 음성을 사용하는 것은 실제 상황에서 매우 실용적이다. 학습 과정에서 감정별로 독립적으로 감정 모델을 생성하여 사용하는 개별 모델 시스템(separate model)은 감정 데이터가 감정별로 100개 이상 충분히 주어져야만 우수한 성능을 나타내었다.

Ⅳ. 결론

본 논문에서는 기존 감정 인식 시스템의 성능 향상을 위하여 감정 적응을 사용한 감정 학습 방법이 제안되었다. 감정 적응을 위하여 적은 개수의 학습 감정 음성과 감정 적응 방식을 사용하여 감정이 없는 음성 모델로부터 감정 음성 모델이 생성되었다. 화자로부터 학습을 위하여 충분한 감정 음성을 얻는 것은 쉽지 않기 때문에 적은 개수의 감정 음성을 사용하는 것은 실제 상황에서 매우 실용적이다.

제안된 방법은 4가지 감정이 포함된 한국어 데이터베이스를 사용하여 평가되었다. 본 연구에서 제안된 감정이 없이 학습된 모델을 감정 음성을 사용하여 감정 적응하여 감정 모델로 변환하는 방법과 기존 학습 방법의 성능을 비교하였다. 실험 결과에서 제안된 방법은 기존 시스템은 학습 과정에 사용되는 감정 학습 데이터 개수가 적을수록 성능이 급격히 저하되었다. 그러나 감정 적응을 사용하는 제안된 방법은 적은 개수의 감정 음성을 사용하면서 기존 시스템보다 우수한 성능을 나타내었다.

References

[1] Rafael A. Calvo, SSidney D'Mello, "Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications," *IEEE Transactions on the. Affective Computing*, Vol.1, No.1, pp.18-37, 2010. DOI: 10.1109/T-AFFC.2010.1

[2] V. Kostv and S. Fukuda, "Emotion in User Interface, Voice Interaction System," in *Proc.. of the IEEE International Conference on Systems, Cybernetics Representation*, pp.798-803, 2000.

[3] T. Moriyama and S. Oazwa, "Emotion Recognition and Synthesis System on Speech," in *Proc. of the IEEE Intl. Conference on Multimedia Computing and System*, pp.840-844, 1999. DOI: 10.1109/MMCS.1999.779310

[4] L. C. Siva and P. C. Ng, "Bimodal Emotion Recognition," *in Proc. of the 4th Intl. Conference on Automatic Face and Gesture Recognition*, pp.332-335, 2000. DOI: 10.1109/AFGR.2000.840655

[5] Kokane Amol T., Ram Mohana Reddy Guddeti, "Multiclass SVM-based Language- Independent Emotion Recognition using Selective Speech Features," *in Proc. of ICACCI*, pp.1069-1073, 2014. DOI: 10.1109/ICACCI.2014.6968337

[6] Rode Snehal Sudhkar, Manjare Chandraprabha Anil, "Analysis of Speech Features for Emotion Detection: A review," *in Proc. of the 2015 International Conference on Computing Communication Control and Automation*, pp.661-664, 2015. DOI: 10.1109/ICCUBEA.2015.135

[7] Y. G. Kim, Y. C. Bae, "Design of Emotion Recognition Model Using fuzzy Logic," *in Proc. of KFIS Spring Conference*, pp.268-282, 2000. DOI: 10.1007/s11042-019-7250-z

[8] K. B. Sim, C. H. Park, "Analyzing the Element of Emotion Recognition from Speech," *Journal of Korean Institute of Intelligent Systems*, Vol.11, No.6, pp.510-515, 2001. DOI: 10.5391/JKIIS.2003.13.1.045

[9] Reynolds, D. A., Quatieri, T. F., Dunn, R. B., “Speaker Verification using Adapted Gaussian Mixture Models,” *Digital Signal Processing,* Vol.10, pp.19-41, 2000. DOI: 10.1006/dspr.1999.0361
[10] Natalia Tomashenko1,2 and Yannick Esteve, “Evaluation of Feature-Space Speaker Adaptation for End-to-End Acoustic Models,” *in Proc. of the Eleventh International Conference on Language Resources*, pp.3163-3170. 2018.
[11] W. G. Kim, “Speech Emotion Recognition using Feature Selection and Fusion Method,” *Transactions of the Korean Institude of Electrical Engineers,* Vol.66, No.8, pp.1265-1271, 2017.
DOI: 10.5370/KIEE.2017.66.8.1265

BIOGRAPHY

Weon-Goo Kim (Member)

1987 : BS degree in Electronic Engineering, Yonsei University.
1989 : MS degree in Electronic Engineering, Yonsei University.
1994 : PhD degree in Electronic Engineering, Yonsei University.

1994~Present : Professor, Dept. of Electrical Engineering, Kunsan National University
1998~1999 : Consultant, Bell Lab., Lucent Technologies(USA)
2008~2009 Visiting Scholar, Griffith Univerisity