한국음향학회지 제22권 제1호 pp. 23~27 (2003)

연속 은닉 마코프 모델을 이용한 한국어 음성 인식을 위한 효율적 음절 모델링

Effective Syllable Modeling for Korean Speech Recognition Using Continuous HMM

김 봉 완*, 이 용 주**
(Bong-Wan Kim*, Yong-Ju Lee**)

*원광대학교 음성정보기술산업지원센터, **원광대학교 전기, 전자 및 정보공학부
(접수일자: 2002년 7월 15일; 채택일자: 2002년 11월 20일)

최근 연속 음성 인식에서의 성능 향상을 위해 음절을 인식 단위로 사용하고자 하는 노력들이 보고되고 있다. 그러나 음절의 경우 음소에 비해 학습성이 음소에 비해 좋지 않고, 모델의 수가 음소에 비해 매우 많으므로 음절 경계에서의 문맥 종속 모델링이 어렵다는 단점을 갖고 있다. 본 논문에서는 한국어에서의 음절의 학습성을 향상시키기 위한 방법과 음절경계에서의 음소 문맥 종속 음절 모델링을 제안한다. 제안된 방법을 단어 인식 실험에 적용한 결과, 기존의 음절 모델과 비교하여 평균 46.23%의 에러 감소율을 보였다. 우측 음소 종속 음절 모델 (right phone dependent syllable model)의 경우 트라이폰 (triphone) 모델에 비해 16.7%의 에러 감소율을 볼 수 있었다.
핵심용어: 음성인식, 음절 모델, 음향 모델, 인식 단위
투고분야: 음성처리 분야 (2.5)

Recently attempts to use the syllable as the recognition unit to enhance performance in continuous speech recognition have been reported. However, syllables are worse in their trainability than phones and the former have a disadvantage in that context-dependent modeling is difficult across the syllable boundary since the number of models is much larger for syllables than for phones. In this paper, we propose a method to enhance the trainability for the syllables in Korean and phoneme-context dependent syllable modeling across the syllable boundary. An experiment in which the proposed method is applied to word recognition shows average 46.23% error reduction in comparison with the common syllable modeling. The right phone dependent syllable model showed 16.7% error reduction compared with a triphone model.
Keywords: *Speech recognition, Syllable modelling, Acoustic modelling, Recognition unit*
ASK subject classification: *Speech signal processing (2.5)*

I. 서 론

음성 인식을 위한 음향 모델링에서 중요한 문제점 중 하나가 인식 단위의 선정이다. 음성 인식기의 성능에 있어서 인식 단위는 절대적인 영향을 미치기 때문이다. 다양한 단위들이 음성 인식을 위해서 사용 가능하며 음성 인식기의 설계시에 인식기에 알맞은 인식 단위를 선정하는 것이 바람직하다. 현재 음소단위의 경우 민감성을 높이기 위해 좌·우의 음소의 문맥을 고려하는 트라이폰 또는 퀴인폰 (quinphone)과 같은 음소 등이 자주 사용되고 있다.

또한 최근 대화체 연속 음성 인식 시스템의 경우 대화 음성에서 음소가 탈 락되거나 약화되는 경우가 많이 발생함으로써 인식 시스템의 성능을 저하시킴에 따라 음소보다 긴 음절과 같은 단위들을 인식의 단위로 검토하고자 하는 노력들이 보고되고 있다[1,2]. 한국어의 경우에도 특히 한국어의 표기 특성을 잘 나타내는 단위인 음절에 대하여 일부 연구가 진행된 바 있으며[3,4], 선행 연구에 따르면 인식률에 있어서는 음소보다 다소 성능이 낮으나 인식 속도에 있어서는 음소보다 빠른 것으로 보고

책임저자: 김봉완 (bwkim@sitec.or.kr)
570-749 전북 익산시 신용동 344-2
원광대학교 음성정보기술산업지원센터 (SiTEC)
(전화: 063-850-7452; 팩스: 063-850-7454)

된 바 있다[4]. 그러나 이러한 인식률 결과는 제한된 양의 음성 데이터를 학습과 평가에 사용하여 얻어진 결과로 음절의 경우 음소보다 학습성에 있어서 불리한 단점이 그대로 반영된 결과라고 해석될 수 있다. 또한 음절은 음절 경계에서의 문맥 종속 모델링이 어렵다는 단점을 갖고 있다.

따라서 본 논문에서는 한국어의 음성 인식을 위한 음절의 학습성을 향상시키기 위하여 음소 단위의 학습 과정을 거친 후, 음절을 구성하는 음소 모델을 결합하여 음절 모델을 생성하는 방법을 제안한다. 제안된 방법을 단어 인식 실험에 적용한 결과, 처음부터 아무런 사전 정보 없이 더 많은 학습 과정을 거친 음절 모델과 비교하여 평균 46.23%의 에러 감소율을 보였다. 또한 본 논문에서는 음절 경계에서의 문맥 종속 모델링을 위해 음소 문맥 종속 음절 모델링을 제안한다. 제안된 방법을 사용한 우측 음소 종속 음절 모델의 경우 트라이폰 모델에 비해 16.7%의 에러 감소율을 볼 수 있었다.

본 논문의 구성은 다음과 같다. 2절에서는 제안된 학습성이 향상된 음절 모델링에 대하여 기술하고, 3절에서는 문맥 종속 음절 모델링에 대하여 기술한다. 4절에서는 실험 결과에 대하여 언급하고 5절에서 결론을 맺는다.

II. 학습성이 향상된 음절 모델링

한국어의 음절 단위는 영어와 마찬가지로 학습성에 있어서 음소보다 불리하고, 음절간 문맥에 둔감하다는 단점과 음절을 구성하는 음소간 조음 결합을 포함하는 장점을 가지고 있다[4]. 그러나 한국어 음절 단위는 한국어의 특성상 음소보다 학습과 인식에서 음절을 분리해 내기 쉽다는 장점을 갖고 있다. 한국어 음절의 구성은 다음과 같다.

$$\langle C_i \rangle V \langle C_f \rangle$$

여기에서 C_i는 초성 자음을 의미하며 19개의 자음이 나타날 수 있다. V는 중성 모음을 의미하며 22개의 모음이 나타날 수 있다. C_f는 종성 자음을 의미하며 7개의 자음이 나타날 수 있으며 $\langle\rangle$는 생략될 수 있음을 의미한다.

위의 구성에 의해 한국어에서 발생 가능한 음절의 수는 이론적으로 3,520개이나 실제로 쓰이는 수는 음소의 연결에 제약이 있어 이보다 훨씬 적은 것으로 알려져 있으며[5]에 의하면 사전에 수록된 65,973개의 표제어에 나타난 음절의 총 가짓수는 1,453개로 조사된 바 있다. 따라서 영어의 음절에 비해 기본 단위의 수가 현저히 적어 어휘 독립 음성 인식을 위한 단위로서의 장점을 갖고 있다.

그러나 이러한 음절은 음소에 비해 학습용 데이터베이스에서 각 단위별 출현 횟수가 현저히 적어 충분한 학습이 이루어지지 않는 문제점이 있다. 7,000만 어절의 텍스트 코퍼스[6]에서 4,300만 어절을 분석한 결과 음소와 음절의 각 단위별 평균 출현 횟수의 차이는 100배 이상인 것으로 나타났으며, 고빈도 200여 개의 음절을 제외하고 나머지 음절의 경우는 모두 가장 적게 출현한 음소보다도 그 출현 횟수가 적게 나타났다.

따라서 음절의 학습성을 보완할 필요가 있으며, 본 논문에서는 학습성이 우수한 음소 단위의 학습 과정을 거친 후, 음절을 구성하는 음소 모델들을 순차적으로 복사하고 결합하여 음절 모델을 생성하는 방법을 제안한다. 즉 발음 사전에 의한 공유의 경우 각 단위가 실제 결합되지는 않고 공유되는 형태이지만 제안된 방법에서는 음절을 구성하는 각 음소 모델의 파라미터를 음절의 해당 위치로 복사하고 결합하여 새로운 음절 모델을 생성하고, 이 후부터는 생성된 음절을 별도의 학습 과정에 참여시킴으로써 안정되게 학습된 이전 음소의 장점을 수용하면서 이후의 음절 학습 과정에서 음절 내부의 조음 결합을 반영할 수 있다. 위의 방법을 이용하면 비단 문맥 비종속 음소뿐만 아니라 보다 성능이 좋은 문맥 종속 음소를 결합하여 음절을 생성할 수도 있으며, 인식 대상 어휘의 증가에 따른 학습 데이터에서 출현하지 않은 음절 (unseen syllable)도 기본 음소 모델만 갖고 있으면 필요할 때 생성해 낼 수 있다는 장점이 있다.

그러나 이러한 방법을 사용할 경우 생성된 음절 모델을 구성하는 상태의 수가 음절을 구성하는 음소의 수에 비례하게 되는 문제점이 있으므로, 이를 해결하기 위해 본 논문에서는 다음과 같은 상태의 병합 및 분할 방법을 사용하여 생성된 음절 모델의 상태의 수를 조절하도록 하였다.

◆ 상태 병합 방법

1) 최소 상태간 거리 (interstate distance)를 갖는 두 인접한 상태 $i, i+1$을 찾는다.
2) 두 상태의 특성을 반영한 다음과 같은 파라미터를 갖는 새로운 상태 n을 생성한다.

$$\mu_n = average(\mu_i, \mu_{i+1}) \quad (1)$$

$$\sum_n = \max(\sum_i, \sum_{i+1}) \quad (2)$$

$$a_{nn} = \frac{a_{ii} + a_{(i+1)(i+1)}}{2} \quad (3)$$

$$a_{n(i+2)} = 1 - a_{nn} \quad (4)$$

$$a_{(i-1)n} = a_{(i-1)i} \quad (5)$$

3) 상태 $i, i+1$을 제거한다.

◆ 상태 분할 방법

1) 최대 상태간 거리를 갖는 두 인접한 상태 $i, i+1$을 찾는다.

2) 새로운 상태 n을 생성하고 그 파라미터를 다음과 같이 설정한다.

$$\mu_n = \mu_i + sign(\mu_{i+1} - \mu_i) \cdot 0.2 \cdot \sigma_i \quad (6)$$

$$\sum_n = \sum_i \quad (7)$$

$$a_{nn} = \frac{a_{ii} + a_{(i+1)(i+1)}}{2} \quad (8)$$

$$a_{n(i+1)} = 1 - a_{nn} \quad (9)$$

$$a_{in} = a_{i(i+1)} \quad (10)$$

$$a_{ij} = 0 \quad (11)$$

위에서 사용된 상태간 거리 $d(i,j)$는 다음과 같다.

- 단일 가우시안의 경우

$$d(i,j) = \frac{1}{S}\sum_{s=1}^{S}\left[\frac{1}{V_s}\sum_{k=1}^{V_s}\frac{(u_{isk}-u_{jsk})^2}{\sigma_{isk}\sigma_{jsk}}\right]^{1/2} \quad (12)$$

- 그 외의 경우

$$d(i,j) = -\frac{1}{S}\sum_{s=1}^{S}\frac{1}{M_s}\sum_{m=1}^{M_s} log[b_{js}(\mu_{ism})] + log[b_{is}(\mu_{jsm})] \quad (13)$$

III. 문맥 종속 음절 모델링

대어휘 음성 인식을 위해서는 문맥 종속 모델링이 필수적이라고 할 수 있다. 그러나 음소의 경우 기본 단위의 수가 적으므로 이러한 문맥 종속 모델링이 어렵지 않게 수행될 수 있으나, 한국어 음절의 경우 약 1,500여 개에 달하므로 문맥 종속 모델링을 수행하기 매우 어렵다. 극단적인 경우 1,500여개의 음절에 대하여 앞, 뒤의 문맥을 고려하여 문맥 종속 모델링의 경우 $3,375 \times 10^6$ (=1,500 × 1,500 × 1,500) 개의 모델이 필요하며, 앞 또는 뒤의 한가지 문맥만을 고려하여 문맥 종속 모델링의 경우에도 $2,250 \times 10^3$ (=1,500 × 1,500)개의 모델이 필요하여 현재의 컴퓨터 시스템으로는 실제로 구현하기가 불가능에 가깝다고 볼 수 있다.

따라서 이러한 음절의 문맥 종속 모델링의 어려움을 극복하기 위하여 본 논문에서는 음소 문맥 종속 음절 모델링을 제안한다. 즉, 문맥 종속 모델링을 할 때 앞 또는 뒤의 음절 전체를 문맥의 구성 요소로 사용하는 것이 아니라 앞 음절의 경우 음절을 구성하고 있는 가장 나중 음소를 문맥 구성 요소로 사용하고, 뒤 음절의 경우 음절을 구성하고 있는 가장 첫 음소를 문맥 구성 요소로 사용하는 것이다.

예를 들어 그림 1과 같이 앞과 뒤의 문맥을 고려하는 좌, 우측 음소 종속 음절 모델링의 경우 약 $1,783 \times 10^3$ (모음 또는 종성 자음 29 × 1,500 × 초성자음 또는 모음 41)개의 모델을 통하여 모델링이 가능하다.

또한 앞의 문맥만을 고려하는 좌측 음소 종속 음절 (left phone dependent syllable) 모델의 경우 43,500 (= 모음 또는 종성 자음 29 × 1,500)개의 모델, 뒤의 문맥만을 고려하는 우측 음소 종속 음절 모델의 경우 61,500 (=1,500 × 초성자음 또는 모음 41) 개의 모델을 통하여 문맥 모델링이 가능하다. 그러나 이와 같은 모델의 수는 음소가 모두 출현한다는 것을 가정한 극단적인 경우로, 실제에 있어서는 음소의 연결 제약으로 인해 그 수가 이보다는 훨씬 적게 나타나리라고 쉽게 예측할 수 있다. 실제 4,300만 어절 텍스트 코퍼스에서 나타난 고빈도 10만 어절에 대한 분석에서도 우측 음소 종속 음절 모델의 경우 총 12,000여 개의 모델만이 출현하였다.

이러한 음소 문맥 종속 음절 모델링은 학습과 인식을 위해 복잡한 알고리즘을 고안하거나 기존 알고리즘의 수정 없이 쉽게 적용할 수 있는 장점을 갖고 있다고 할 수 있다. 본 논문에서는 좌측 음소 종속 음절 모델을 사용하여 실험하였다.

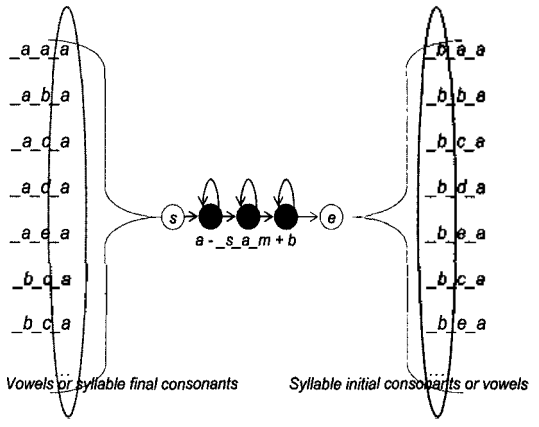

그림 1. 좌우 음소 종속 음절 모델링
Fig. 1. LR phone dependent syllable modeling.

IV. 실험 결과

4.1. 음성 데이터베이스

인식 실험에 사용한 음성 데이터 베이스는 PBW 452어절 데이터 베이스[7]로, 표준어를 사용하는 70명의 화자가 방음실에서 452어절의 음소 균형 단어를 2회 발성한 데이터로 구성되어 있다. 70명의 화자 중 남성 화자의 수는 38명, 여성 화자의 수는 32명으로 구성되어 있으며 연령 분포는 20대와 30대가 주를 이루고 있다. 음성 신호의 수집에 사용된 마이크로폰은 Sennheiser HMD 224X headset을 사용하였으며 발성한 신호는 16 kHz, 16 bit linear PCM으로 A/D 변환하였다.

총 70명의 화자에 의해 수집된 음성 자료는 총 36.8시간 분량이며, 학습에 50명 (남성 화자 25명, 여성 화자 25명)의 데이터를 사용하였으며 그 양은 26.56시간 분량이다. 평가에 사용한 음성 데이터는 20명 (남성 화자 13명, 여성 화자 7명)으로 구성되어 있으며 그 양은 10.24시간 분량이다.

음운 균형 단어 (PBW) 데이터베이스의 경우 총 451개의 음절이 출현하였으며, 음소와 음절의 1회의 학습 과정으로 학습될 수 있는 학습량을 다음 표 1에 나타내었다. 학습량에 있어서 음소 모델을 사용한 경우가 음절 모델을 사용한 경우 보다 평균 21배 이상의 학습량을 보임을 알 수 있다.

표 1. PBW DB에서의 음소, 음절 출현 횟수
Table 1. The number of occurrence of phoneme and syllable models in PBW DB.

Recognition Unit	Number of models	Average number of occurrence	Standard deviation
Phoneme	42	7261.905	4755.944
Syllable	451	340.576	461.754

4.2. 실험 및 결과

학습 및 인식에 사용된 특징 벡터는 25 msec의 창(window)을 10 msec씩 전진시키면서 12차의 MFCC와 1차의 에너지, 이들에 대한 차분 파라미터 (delta parameters), 차분에 대한 차분 파라미터 (acceleration parameters) 등 총 39차의 특징 벡터를 사용하였으며, 음향 모델은 단순 좌우 위상을 갖는 연속 은닉 마코프 모델로 모델링하였다.

먼저 기본 음소 단위의 결합에 의한 음절 생성 후 상태의 수 조절 방법을 적용하기 위해 각 모델별 3개의 상태를 갖는 기본 음소 단위 시스템을 구성하였다. 기본 음소 단위 시스템의 성능은 3회의 학습과정을 거친 시스템이 70.25%, 5회의 학습과정을 거친 시스템이 75.82%의 인식률을 나타냈다.

제안된 음절 생성 방법의 성능을 검증하기 위한 비교 시스템으로 플랫 스타트 (flat start) 방법을 이용하여 음절 모델을 구성하고 5회의 학습 과정을 거치도록 하였다. 또한 음절 모델별 상태의 수가 3~9인 시스템들을 각각 구성하여 성능 변화를 살펴보았다. 제안된 시스템은 3회의 학습과정을 거친 음소모델을 이용하여 음절 모델을 생

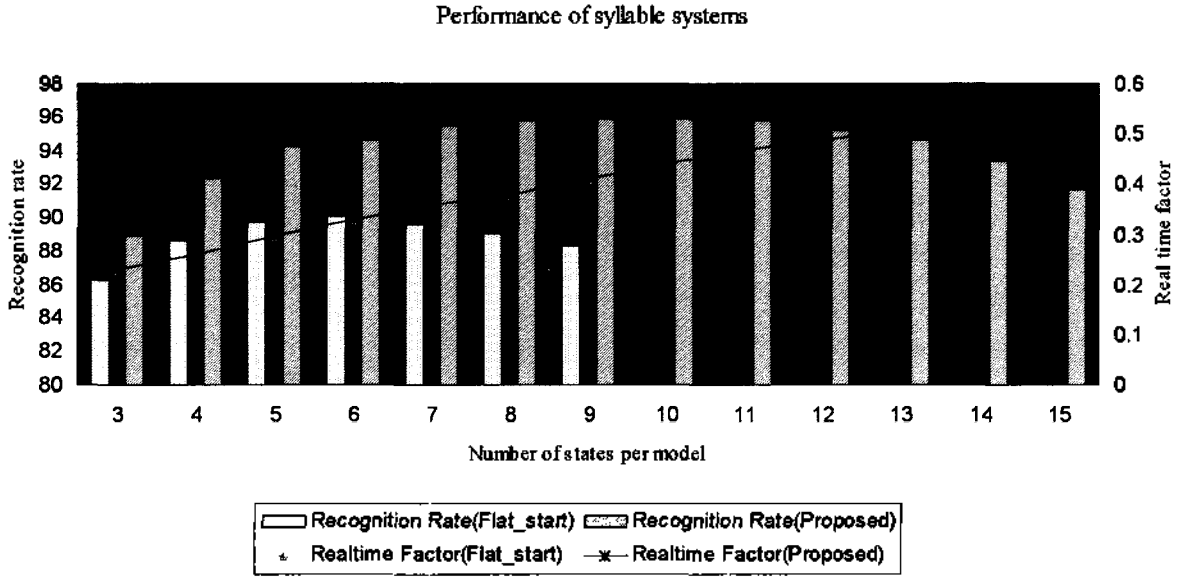

그림 2. 음절 시스템들의 성능 비교
Fig. 2. Performance of syllable systems.

표 2. right phone dependent syllable 기반 시스템, triphone 기반 시스템의 단어 인식 성능 비교
Table 2. Performance of right phone dependent syllable system and triphone system.

System	Number of models	Number of States	Recognition Rate	Real time factor
Triphone	2134	6405	99.28	0.50
Right phone dependent syllable	1090	5450	99.40	0.46

성한 후 제안된 방법을 이용하여 모델별 상태의 수가 3~15인 각각의 시스템을 구성하고 2회의 학습 과정만을 거치도록 하였다.

비교 시스템들과 제안된 시스템들의 성능은 그림 2에 나타나 있으며, 모델별 상태의 수가 3~9까지를 비교해 보면 제안된 시스템이 에러 감소율에 있어 평균 46.23%를 나타내고 있어 제안된 방법이 유효함을 알 수 있다.

또한 음소 문맥 종속 음절 모델링의 유효성을 검증하기 위하여 모델별 5개의 상태를 갖는 우측 음소 종속 음절 모델 시스템을 구성하고 트라이폰과 성능을 비교한 결과는 다음 표 2와 같다. 결과를 보면 오히려 적은 상태의 수에도 불구하고 트라이폰 모델에 비해 16.7%의 에러 감소율을 볼 수 있으며 인식 속도에 있어서도 더 빠른 인식 시간을 보이고 있다.

V. 결 론

본 논문에서는 한국어의 음성 인식을 위한 음절의 학습 성을 향상시키기 위하여 음소 단위의 학습 과정을 거친 후, 음절을 구성하는 음소 모델을 결합하여 음절 모델을 생성하는 방법과 음절 경계에서의 문맥 종속 모델링을 위해 음소 문맥 종속 음절 모델링을 제안한다. 제안된 음절 생성 방법을 단어 인식 실험에 적용한 결과 평균 46.23%의 에러 감소율을 보였으며, 우측 음소 종속 음절 모델의 경우 단어 인식 실험에서 트라이폰 모델에 비해 16.7%의 에러 감소율을 볼 수 있었다. 향후 연구 과제로는 음절 생성시 상태의 수를 자동으로 결정하기 위한 음절의 길이를 모델링할 수 있는 방안에 대한 연구와, 학습 데이터에 출현하지 않은 음절을 생성할 때 단순히 기존 음소를 복사하여 결합하는 것이 아니라 생성된 후 학습에 의해 조음 결합이 반영된 유사 음절들간의 적절한 상태 결합에 의한 음절의 생성 방법에 대한 연구를 진행하고자 한다.

감사의 글

본 연구의 일부는 원광대학교 2000년 교비연구의 지원에 의한 것임.

참 고 문 헌

1. Aravind Ganapathiraju, Jonathan Hamaker, Joseph Picone, Mark Ordowski, George R. Doddington, "Syllable-based large vocabulary continuous speech recognition," *IEEE Trans. on Speech and Audio Processing*, **9** (4), 358-366, 2001
2. H. Bourlard, H. Hermansky, and N. Morgan, "Copernicus and the ASR challenge-waiting for kepler," *Proc. DARPA Speech Recognition Workshop*, 157-162, 1996.
3. 이영호, 정궁, "음절을 기반으로 한 한국어 음성인식," 전자공학회 논문집, **31-B** (1), 11-22, 1994.
4. 김유진, 김회린, 정재호, "인식 단위로서의 한국어 음절에 대한 연구," 한국음향학회지, 16 (3), 64-72, 1997.
5. KBS, 표준 한국어발음대사전, 어문각, 1993.
6. K.-S. Choi, KAIST 언어자원 2001년도판, 과학기술부 핵심 소프트웨어 과제 결과물 1995-2000 (http://kibs.kaist.ac.kr).
7. B.-W. Kim, S.-T. Kim, T.-W. Kim, Y.-I. Kim, and Y.-J. Lee, "Design and construction of korean speech database for common use," *Proc. ICSP 97*, 2, 1997.

저자 약력

● **김 봉 완 (Bong-Wan Kim)**

1995년: 원광대학교 컴퓨터공학과 (공학사)
1997년: 원광대학교 컴퓨터공학과 (공학석사)
2002년: 원광대학교 컴퓨터공학과 (공학박사)
2001년~현재: 원광대학교 음성정보기술산업지원 센터 선임연구원
※ 주관심분야: 음성인식, 음향 모델링, 음성 DB

● **이 용 주 (Yong-Ju Lee)**

1976년: 고려대학교 전자공학과 (공학사)
1987년: 고려대학교 전자공학과 (공학석사)
1992년: 고려대학교 전자공학과 (공학박사)
1976년~1980년: 공군 제7항로보안단 통신전자 장교
1980년~1994년: 한국전자통신연구소 자동통역 연구실 실장 (책임연구원)
1994년~현재: 원광대학교 전기 전자 및 정보공학부 부교수
2001년~현재: 원광대학교 음성정보기술산업지원센터 센터장
※ 주관심분야: 음성인식, 음성합성, 음성정보기술, 음성 DB