음성학적 지식 기반 변이음 모델을 이용한 가변 어휘 단어 인식기

Variable Vocabulary Word Recognizer using Phonetic Knowledge-based Allophone Model

김 회 린*, 이 항 섭*
(Hoi-Rin Kim*, Hang-Seop Lee*)

※이 연구는 정보통신부의 지원으로 이루어진 결과물 입니다.

요 약

본 논문에서는 훈련용 음성 데이터와 무관한 임의의 새로운 어휘를 인식해 낼 수 있는 가변 어휘 단어 인식기 개발에 대하여 기술한다. 가변 어휘 단어 인식기를 구현하기 위해서는, 인식 대상이 될 새로운 어휘를 즉시 발음 사전으로 변환시키는 on-line 발음 사전 생성기가 필요하고, 발음 사전 출력을 가지고 각 단어를 모델링할 수 있는 신뢰성 있는 음소 및 변이음 모델이 필요하다. 이와 같은 신뢰성 있는 음소 및 변이음 모델을 생성시키기 위하여 본 연구에서는, 각 음소와 전후 음소들의 음성학적 자질을 고려하여 3 음소열을 집단화(clustering)하여 변이음을 정의하고 이를 당 연구실이 보유하고 있는 POW(Phonetically Optimized Words) 3,848개 단어에 적용하여 1,548개의 변이음 모델을 생성시켰다. 이를 토대로 가변 어휘 단어 인식기를 구현하고 이를 POW 3,848 DB, PBW 445 DB 및 호텔 예약용 244 단어 DB 등에 적용하여 그 성능을 평가하였다. 평가 결과, POW DB에 대해서는 79.6%, PBW DB에 대해서는 445 단어 사전의 경우 79.4%, 100 단어 사전의 경우 88.9%의 성능을 보여 주었고, 호텔 예약 DB에 대해서는 71.4%의 성능을 보여 주었다.

ABSTRACT

In this paper, we propose a variable vocabulary word recognizer that is able to recognize new words not exist in training data. For the variable vocabulary word recognizer, we must have an on-line lexicon generator to transform new candidate words to the corresponding pronunciation sequences of phones without any large lexicon table. And, we also must make reliable models of all the phones and allophones, so as to precisely model the new words with the lexicon generator outputs. In order to model the phones and allophones reliably, we define Korean allophones by triphone clustering based on phonetic knowledge of preceding and succeeding phones of each phone. Using the clustering method, we generated 1,548 allophones with POW (Phonetically Optimized Words) 3,848 word DB. We evaluated the proposed word recognizer with POW 3,848 DB, PBW (Phonetically Balanced Words) 445 DB, and 244 word DB in hotel reservation task. Experimental results showed word recognition accuracy of 79.6% for the POW DB corresponding to vocabulary-dependent case, 79.4% in case of 445 word lexicon and 88.9% in case of 100 word lexicon for the PBW DB, and 71.4% for the hotel reservation DB corresponding to vocabulary-independent case.

Ⅰ. 서 론

음성 인식 시스템을 구현할 때, 우리는 보통 구현하고자 하는 인식기가 인식할 대상 어휘를 미리 선정하고, 이 어휘들에 대해서 음성 데이터 베이스를 수집한다. 그리고 이 음성 DB를 사용하여 인식할 단어나 음소의 모델을 훈련하게 된다. 이렇게 하여 구현한 인식기는 처음에 정의한 어휘에 적합하도록 단어나 음소 모델이 훈련되므로 미리 정의된 어휘에 대해서는 인식률이 높게 동작하지만 새로운 어휘를 인식할 필요가 있을 때는 처음부터 다시 모델을 훈련해야 하는 불편함이 존재한다.

또한 최초의 훈련에 사용한 음성 DB에 모든 한국어 음소가 포함되어 있지 않았거나, 모두 포함하고 있더라도 음소 환경이 충분히 다양하지 않으면, 이를 이용하여 어

*한국전자통신연구소 음성언어연구실
접수일자: 1996년 10월 16일

휘에 무관한 음성 인식기를 만들 때 인식 성능의 저하를 피할 수 없게 된다. 그러므로, 인식할 대상 어휘의 변경 시, 새로운 어휘에 대한 훈련용 음성을 새로 수집하지 않고도 성능이 우수한 음소 모델을 얻기 위해서는 처음에 사용하는 훈련용 음성의 음소 환경적 특성이 매우 중요하게 된다.

본 논문에서는 현재 ETRI에서 개발 중인 가변 어휘 단어 인식기의 1단계 구현 과정을 소개하고, 이 인식기의 성능을 어휘 종속 및 어휘 독립에 대하여 평가한 결과를 기술한다.

Ⅱ. 가변 어휘 단어 인식기 개요

가변 어휘 단어 인식기는 그림 1에서와 같은 구조를 갖는다. 이 인식기는 기존의 인식기와 달리 인식 대상으로 하는 단어 목록이 매 음성 입력 마다 바뀌어도 인식할 어휘에 대한 음성 훈련 과정을 새로 수행하지 않고 단지 발음 사전 만을 새로 교체하여 단어 모델들을 재구성 하므로 이론적으로 무제한의 임의의 단어를 주어진 단어 목록 내에서 인식할 수 있게 된다. 이러한 단어 인식기를 구현하려면, 우선 한국어에 존재하는 모든 음소를 충분한 음소 환경에서 정확히 모델링 해야 한다. 이렇게 하기 위해서는 먼저 각 음소를 정확히 모델링하기 위하여 훈련 데이터를 다양한 음소 환경하에서 수집해야 하며, 또 이를 음소 모델에 적절히 반영시키기 위하여 이러한 다양성을 포용할 수 있는 음소 모델 구조를 가져야 한다. 이러한 조건을 충족시키기 위하여 본 연구에서는 훈련용 음성 데이터로써 당 연구실이 제안한 POW 3,848 단어 목록[1]을 사용하여 다수의 화자로부터 음성을 수집하여 사용하고, 음소의 다양성을 모델 구조에 반영하기 위하여 음소만이 아닌 변이음까지의 상세 모델링을 함으로써 각 모델의 정확도를 향상시켰다.

마지막으로, 매 어휘 변경 요구가 사용자로부터 입력되면 이를 즉시 단어 목록에 반영하고 이로부터 각 단어의 발음 사전이 사전 생성기를 통하여 출력될 수 있도록

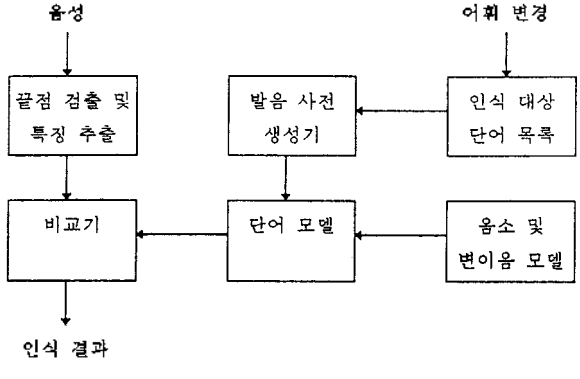

그림 1. 가변 어휘 단어 인식기 구조
Fig 1. Block diagram of variable vocabulary word recognizer

하는 기능이 필요하다. 이 기능도 당 연구실이 보유하고 있는 사전 생성기를 이용하여 구현하였다.

Ⅲ. 음성학적 지식 기반 변이음 정의

음소를 기본 단위로 한 음성 인식 시스템의 성능 향상을 위해서는 각 음소를 다양한 주변 음소 환경 하에서 정확히 모델링하는 것이 중요하다. 이를 위해 사용하는 가장 일반적인 접근 방법은 각 음소의 전후 음소에 따라 각 음소를 달리 모델링하는 3 음소열 (triphone) 모델링 기법이다. 이 기법을 사용하면 각 언어를 대표하는 모든 3 음소열의 총 가짓수가 급격히 증가(최소 1만개 이상)하게 된다. 이와 같은 3 음소열 수의 급격한 증가는 실제 인식 시스템을 구성할 때 메모리나 계산량의 급격한 증가 외에도 모든 3 음소열을 정확히 모델링하기에 충분한 훈련 데이터를 확보하는 것이 거의 불가능 하게 된다.

이에 대한 대안으로 제시되는 방법들로 대표적인 것으로 음성학적 지식 기반 변이음 추출법[2]과 정보 이론적 3 음소열 집단화 방법[3]이 있다. 정보 이론적 집단화 방법은 각 3 음소열 모델의 유사도를 모델 파라메터 영역에서 계산하여 이를 기준으로 집단화 하여 3 음소열을 보다 적은 수의 문맥 종속형 음소 모델들로 대표 시킨다. 이 방법은 인식하고자 하는 어휘가 미리 정해져 있을 때, 이 어휘 내의 단어 집합에 대하여 최적의 문맥 종속형 음소 모델을 생성 시킬 수 있지만, 가변 어휘 인식기에서와 같이 인식 대상 어휘가 훈련 데이터와 무관할 경우에는 새롭게 출현하는 3 음소열을 정의된 문맥 종속형 음소 모델에 적절하게 대응 시키기가 어렵게 된다. 한편, 지식 기반 변이음 추출법은 적용되는 음성학적 지식이 전문가들 사이에 다소 이견이 있을 수 있음에도 불구하고 일단 적용된 규칙은 새로운 3 음소열에 대하여 적절한 변이음 모델을 대응 시키게 된다. 이러한 이유로 본 연구에서는 한국어의 음성학적 지식을 기반으로 변이음을 정의하였다.

음성학적 지식을 기반으로 한 변이음 정의는 당 연구실이 개발한 한국어 대화체 음성 인식 시스템[4]에서 사용한 규칙[5]을 기본으로 하되, 이를 일부 단순화하여 재정의 하였다. 이와 같이 정의된 변이음 추출법에 따라 생성될 수 있는 변이음의 종류는 이론적으로 2,551개가 가능 하다. 한편 변이음 추출에 사용된 텍스트 DB는 POW 3848 단어 DB를 이용하였다. 변이음 추출 과정을 보다 상세히 설명하면 다음과 같다. 먼저, 한국어에 나타나는 모든 음소를 한 개의 묵음 모델을 포함하는 40개의 음소로 대표 시켰다. 이를 기준으로 POW 3,848개 단어 내에 나타나는 모든 3 음소열을 구한 결과 그 수는 모두 9,394개였다. 여기에 제안된 변이음 추출법을 적용하여 총 1,877개의 변이음을 구하고, 이중에서 그 발생 빈도가 극히 적은 변이음을 삭제하여 최종적으로 1,548개의 변이음을 추출하였다. 결국 추출된 변이음의 종류는 이론적으로 가능한 모든 변이음의 60.7%를 포함하고 있다. 여

표 1. 각 음소별 변이음 종류
Table 1. Number of allophones for each phoneme

음소	이론적으로 가능한 변이음 종류	POW DB내의 변이음 종류	음소	이론적으로 가능한 변이음 종류	POW DB내의 변이음 종류
ㄱ	49	40	ㅑ	81	26
ㄲ	49	30	ㅓ	81	73
ㅋ	49	29	ㅕ	81	59
ㄴ	49	44	ㅗ	81	72
ㄷ	49	36	ㅛ	81	47
ㄸ	49	26	ㅜ	81	74
ㅌ	49	32	ㅠ	81	38
ㄹ	49	44	ㅐ	81	61
ㅁ	49	43	ㅒ	81	1
ㅂ	49	37	ㅔ	81	55
ㅃ	49	16	ㅖ	81	28
ㅍ	49	30	ㅡ	81	56
ㅅ	49	33	ㅣ	81	72
ㅆ	49	28	ㅚ	81	25
ㅇ	49	30	ㅘ	81	42
ㅈ	49	32	ㅝ	81	20
ㅉ	49	26	ㅙ, ㅞ	81	51
ㅊ	49	29	ㅟ	81	50
ㅎ	49	28	ㅢ	81	10
나	81	75	합계	2,551	1,548

기에서 누락된 변이음들은 실제로 발생하지 않는 변이음들이거나, 발생하더라도 그 발생 빈도가 극히 적은 것들이다. 그 이유는 POW를 추출할 때 사용한 알고리즘에 텍스트 모집단에서 발생하는 음소 환경의 빈도가 고려되었기 때문에 POW내에 존재하지 않는 변이음 그룹은 실제 생활에서 발성 빈도가 매우 적다고 볼 수 있기 때문이다. 표 1에 각 음소별로 제안된 변이음 추출법을 적용한 결과 이론적으로 가능한 변이음의 종류 및 실제로 추출된 변이음의 종류를 기술하였다.

여기에서 정의한 1,548개의 변이음으로 대응시킬 수 없는 3 음소열이 새로운 발음 사전에 입력될 경우에는 이를 문맥 독립형 음소 모델로 대응시켜서 임의의 3 음소열에 대해서도 음소 및 변이음의 연결로 표현되는 정밀한 단어 모델을 구성할 수 있게 된다. 결국 본 시스템에서 사용되는 음소 및 변이음의 총 가짓수는 1,588개가 된다.

Ⅳ. 음소 및 변이음 모델 훈련

1. 훈련용 음성 DB

가변 어휘 단어 인식 시스템에 사용될 음소 및 변이음 모델의 훈련용 음성 데이터는 POW 3848 DB로써 이는 다음과 같이 구성되어 있다.

POW 3848 DB는 어휘가 총 3,848개로 구성되어 있으며, 이를 8명이 481개씩 나누어 발성한 것을 1개의 set으로 하였다. 이러한 set이 남성음에 대하여 5 set (총 40명), 여성음이 5 set (총 40명)이 있어서 모두 합하면 10 set (약 38,480개 단어)이 된다. 총 10 set 중 남녀 각 4 set 씩 모두 8 set을 각종 모델의 훈련에 사용하며, 이 중에서 남성음 3 set과 여성음 2 set은 수작업으로 음소 경계가 labeling 되어 있다.

이와 같은 POW 전체 DB 중 음소 모델(문맥 독립)의 훈련을 위해 사용한 것은 수작업으로 labeling되어 있는 남성음 3 set과 여성음 2 set (약 19,240 단어) 이다. 변이음 모델(문맥 종속)의 훈련을 위해 사용한 것은 수작업으로 labeling되어 있는 남성음 3 set과 여성음 2 set을 포함하는 남성음 4 set, 여성음 4 set (약 30,784 단어)이다.

이 음성 DB는 비교적 조용한 녹음실에서 수집되었고, 16 kHz, 16 bit로 양자화된 후 이를 끝점 검출기에 통과시켜 각 단어의 시작점 및 끝점을 표시하였다. 이때, 각 단어의 앞뒤에는 약 300 msec 정도의 묵음 구간이 존재하도록 하였다.

2. 특징 추출 및 훈련

특징 벡터 추출 과정은 다음과 같다. 먼저, 10 msec

(160 samples) 마다 256 point FFT를 수행하고, 이로부터 PLP (perceptually linear prediction) 특징 벡터를 구한다. 구해진 특징 벡터로부터 dynamic feature를 구하기 위해 FIR filter를 사용하여 first-order dynamic feature를 얻고, 이 두 가지 벡터를 연결한 26차 벡터에 mean-subtraction을 이용한 정규화를 거쳐 최종적인 26차 특징 벡터를 구한다.

정의된 문맥 독립형 음소 40개 모델(묵음 모델 포함)의 훈련은 앞서 기술한 바와 같이 labeling된 POW 5 set을 가지고 수행한다. 각 음소는 해당 음소 당 50개의 codeword를 가지는 SCHMM으로 모델링 되며, 모델의 구조는 3-state left-to-right (no skip path) model로 정의되어 있다. 또한, 훈련시에 각 단어의 전후에 additive silence model을 사용하였으며, 수작업으로 labeling되어 있는 음소 경계 정보를 그대로 이용하여 각 음소의 codebook 및 distribution을 훈련하였다.

문맥 종속형 변이음 모델의 훈련 과정은 다음과 같다. 먼저, 변이음의 정의는 앞서 기술한 바와 같이 한국어 음성학의 지식을 기반으로 1,548개의 변이음 모델을 정의한다. 정의된 각각의 변이음도 위의 음소 모델과 같은 HMM 구조를 가지지만, 차이점은 각 음소 당 50개의 codeword를 가지는 대신에, 각 음소의 state 당 50개의 codeword를 가진다는 점이다. 이렇게 함으로써 주어진 데이터의 양에 적절하면서도 보다 정밀한 변이음 모델링을 가능케 할 수 있게 된다.[6]

변이음 모델의 훈련은 초기 HMM parameter로서 음소 모델의 parameter를 사용하고, 이 초기 모델과 8 set의 POW DB를 이용하여 bootstrapping 방식으로 iteration 및 codebook 초기화 과정을 반복하여 최종적인 변이음 모델을 구한다.

V. 성능 평가

1. 평가용 음성 DB

구현된 가변 어휘 단어 인식기의 성능을 평가하기 위하여 3가지 종류의 DB를 사용하였다. 첫번째로, 어휘 종속의 성능을 평가하기 위하여 POW DB 총 10 set 중 훈련에 사용되지 않은 남성 및 여성음 각 1 set을 이용하였다. 어휘 독립 인식 실험용 DB는 POW와 관계없는 새로운 단어 set으로서 PBW 445 DB와 호텔 예약용 244 단어 음성 DB를 이용하였다. PBW DB는 다시 445개 어휘 전체를 사전으로 사용하는 경우와 이중 100개 만을 사전으로 사용하는 경우로 나누어 실험하였다. 평가용 음성 DB의 보다 상세한 내용이 표 2에 기술되어 있다. 또한 모든 평가용 DB의 녹음, 양자화 및 끝점 검출은 훈련용 DB와 거의 동일하다.

표 2. 평가용 음성 DB 목록
Table 2. List of speech database for evaluation

DB 종류	화자 수	각 단어 당 음성 샘플 수	총 단어 수
POW 3848 DB	남자 8명 여자 8명	2개	3,848 * 2 = 7,696개
PBW 445 DB	남자 22명 여자 19명	41개	445 * 41 = 18,245개
호텔 예약용 244 단어 DB	남자 9명	9개	244 * 9 = 2,196개

2. 실험 결과 및 분석

표 3에 제안된 가변 어휘 단어 인식기의 화자 독립 인식 성능이 기술되어 있다. 단어 인식 시에 Viterbi beam search를 사용하였으며, 이때 beam threshold는 2.5로 하였다. 이 값은 여러 가지 다른 값에 대하여 실험해 본 결과이다. 이 표에서 "Iter. = 1"은 변이음 모델 훈련 시 bootstrapping 훈련의 HMM 파라메터 재추정 단계가 첫번째 임을 의미한다. 또한, PBW 445 DB를 이용하여 평가할 때 사전 크기가 100인 경우는, 445개 어휘 중 100개씩의 어휘를 4 종류 추출하고 이에 대한 인식 성능을 각각 구한 후, 이를 평균한 것이다.

먼저 어휘 종속의 실험 결과를 보면, 음소에 비해 변이음 모델을 사용할 경우 인식률이 크게 향상되었음을 보여준다. 또한 인식률이 80%를 넘지 못한 이유는 검색 사전이 3,848개의 어휘로 구성되어 있고, 이들이 1 음절에

표 3. 가변 어휘 단어 인식기의 화자 독립 인식 성능 평가 결과
Table 3. Performance evaluation results for speaker-independent recognition of variable vocabulary word recognizer

실험 조건	DB 종류	사전 크기	단어 인식률(%)			
			음소 모델만 이용 (40개)	음소 및 변이음 모델 이용(1,588개)		
				Iter. = 1	Iter. = 2	Iter. = 3
어휘 종속	POW	3,848	71.4	67.1	79.6	78.0
어휘 독립	PBW	445	76.1	69.5	78.4	79.4
		100	87.9	84.2	88.0	88.9
	호텔예약	244	68.4	63.9	71.4	66.5

서 9 음절까지 골고루 분포하고 있어서 적은 수의 음절로 구성된 경우 오인식률을 증가시켰기 때문인 것으로 분석된다.

어휘 독립의 경우, 변이음 모델을 사용할 때의 오인식률이 음소 모델만을 사용할 때에 비해서 10% 내외 성능만 감소하였다. 이것은 훈련과 실험용 DB 어휘의 상이성이 인식률 향상에 큰 저해 요인임을 보여주고, 변이음의 종류가 가변 어휘 인식기의 경우 너무 많지 않은 것이 바람직하다는 것을 반증한다. 또한 사전 크기가 어휘 종속에 비하여 크게 작음에도 불구하고 전반적으로 인식률이 낮았는데, 이것도 앞서 기술한 어휘의 상이성 때문이다. 마지막으로 호텔 예약 DB의 경우 사전 크기가 244개 정도임에도 불구하고 다른 어휘 독립 실험에 비하여 상대적으로 저조한 성능을 보여 주었는데, 이것은 이 DB의 어휘 내에 숫자음 및 영어 알파벳 단어가 약 2/3 정도나 되어 유사한 발성이 매우 많기 때문이다. 하지만 이 경우에도 변이음 모델을 사용할 때 적절한 인식률의 향상이 있었다.

이러한 결과로 볼 때, 가변 어휘 단어 인식기의 성능을 보다 개선시키기 위해서는 적절한 수의 변이음 종류를 사용하는 것과 함께, 강력한 검색 어휘 지식의 이용이 필요함을 결론 지을 수 있다.[7]

Ⅵ. 결 론

본 논문에서는 가변 어휘 단어 인식기를 구현하는 방법을 기술하고, 구현 시 음성학적 지식에 기반을 둔 변이음 모델을 사용하는 것이 성능 향상에 기여함을 보였다. 성능 평가 결과 음소 모델만으로도 비교적 높은 인식률을 보인 것은, 비록 음소 모델만의 훈련 시에도 인식시에 새로운 어휘를 인식 시키려면 훈련시 가능한 한 다양한 음소 환경을 가진 데이터를 사용해야 함을 보여주었다. 또한, 변이음 모델을 사용할 경우의 인식률이 음소 모델만 사용했을 경우보다 성능 향상이 아주 크지 않았는데, 이것은 가변 어휘 인식 시스템을 구현할 때 여기에서와 같은 단순한 방법 이외의 보다 강력한 검색 어휘 지식의 이용이 필요하게 됨을 보여준다.

이상과 같은 결과로부터 당 연구실에서는 보다 실용적인 인식 시스템의 구현을 위해 첫 단계로 PC에 이 인식 시스템을 구현[8]하였고, 이를 바탕으로 가변 어휘 단어 인식기의 지속적인 성능 향상 및 이의 응용 시스템 개발을 추진할 계획이다.

참 고 문 헌

1. Yeonja Lim and Youngjik Lee, "Implementation of the POW (Phonetically Optimized Words) algorithm for speech database," *Proc. of ICASSP*, pp. 89-91, 1995.
2. L. Deng, M. Lennig, V.N. Gupta, P. Mermelstein, "Modeling acoustic-phonetic detail in an HMM-based large vocabulary speech recognizer," *Proc. of ICASSP*, pp. 509-512, 1988.
3. Kai-Fu Lee, *Automatic Speech Recognition*, Kluwer Academic Publisher, pp. 103-106, 1989.
4. 이항섭, 박준, 권오욱, "한국어 대화체 인식 시스템의 구현," *제13회 음성통신 및 신호처리 워크샵(KSCSP'96)* 논문집, 13권, 1호, pp. 145-148, 1996.
5. 서영주, 성철재, 이정철, 한민수, 이영직, "음성학적 지식에 기반한 한국어 변이음 집단화 수형도의 구현," *제13회 음성통신 및 신호처리 워크샵(KSCSP'96)* 논문집, 13권, 1호, pp. 344-347, 1996.
6. M. Hwang and X. Huang, "Subphonetic modeling with Markov states-SENONE," *Proc. of ICASSP*, pp. I-33-36, 1992.
7. 김회린, 이항섭, "POW 3848 단어 인식기 구현 및 어휘 독립 실험," *제13회 음성통신 및 신호처리 워크샵(KSCSP'96)* 논문집, 13권, 1호, pp. 127-130, 1996.
8. 이항섭, 김회린, 이정철, 김상훈, "PC에서의 어휘 독립 및 화자 독립 단어 인식기 구현," *제13회 음성통신 및 신호처리 워크샵(KSCSP'96)* 논문집, 13권, 1호, pp. 192-194, 1996.

▲김 회 린(Hoi-Rin Kim)

1984년 2월 : 한양대학교 전자공학과 졸업(학사)
1987년 2월 : 한국과학기술원 전기 및 전자공학과 졸업(석사)
1992년 2월 : 한국과학기술원 전기 및 전자공학과 졸업(박사)
1994년 6월~1995년 5월 : 일본 음성번역통신연구소(ATR-ITL) 방문연구원
1987년 10월~현재 : 한국전자통신연구소 음성언어연구실 선임연구원
※주관심분야 : 음성인식, 음성언어번역, 음성신호처리

▲이 항 섭(Hang-Seop Lee)

1990년 2월 : 광운대학교 컴퓨터공학과 졸업(학사)
1992년 2월 : 광운대학교 컴퓨터공학과 졸업(석사)
1992년 1월~현재 : 한국전자통신연구소 음성언어연구실 연구원
※주관심분야 : 음성응용시스템, 음성인식, 음성언어번역