확률 발음사전을 이용한 대어휘 연속음성인식

Stochastic Pronunciation Lexicon Model ing for Large Vocabulary Continuous Speech Recognition

윤 성 진*, 최 환 진*, 오 영 환*
(Seong Jin Yun*, Hwan Jin Choi*, Yung Hwan Oh*)

요 약

본 논문에서는 대어휘 연속음성인식을 위한 확률 발음사전 모델에 대해서 제안하였다. 확률 발음 사전은 HMM과 같이 단위음소 상태의 Markov chain으로 이루어져 있으며, 각 음소 상태들은 음소들에 대한 확률 분포 함수로 표현된다. 확률 발음 사전의 생성은 음성자료와 음소 모델을 이용하여 음소 단위의 분할과 인식을 통해서 자동으로 생성되게 된다. 제안된 확률 발음 사전은 단어내 변이와 단어간 변이를 모두 효과적으로 표현할 수 있었으며, 인식 모델과 인식기의 특성을 반영함으로써 전체 인식 시스템의 성능을 보다 높일 수 있었다. 3000 단어 연속음성인식 실험 결과 확률 발음 사전을 사용함으로써 표준 발음 표기를 사용하는 인식 시스템에 비해 단어 오류율은 23.6%, 문장 오류율은 10% 정도를 감소시킬 수 있었다.

ABSTRACT

In this paper, we propose the stochastic pronunciation lexicon model for large vocabulary continuous speech recognition system. We can regard stochastic lexicon as HMM. This HMM is a stochastic finite state automata consisting of a Markov chain of subword states and each subword state in the baseform has a probability distribution of subword units. In this method, an acoustic representation of a word can be derived automatically from sample sentence utterances and subword unit models. Additionally, the stochastic lexicon is further optimized to the subword model and recognizer. From the experimental result on 3000 word continuous speech recognition, the proposed method reduces word error rate by 23.6% and sentence error rate by 10% compare to methods based on standard phonetic representations of words.

I. 서 론

음성은 인간의 언어 소통을 위한 가장 기본적인 수단으로 사용의 간편함이나 의사전달의 효율성 등의 측면에서 다른 방법들에 비해 우수한 것으로 알려져 있다. 이러한 음성이 인간과 기계간의 훌륭한 의사 소통 수단이 되기 위해서 음성인식에 관한 다양한 연구가 필수적이다.

음성인식은 대상으로 하는 음성의 발성 방법에 따라 크게 구분발성 음성과 연속음성으로 나눌 수 있다. 구분발성 음성은 단어 단위로 띄워서 발성하는 방법으로 단어간의 경계가 확실하여 단어간 조음 현상으로 인한 변이가 작아 비교적 높은 인식 성능을 보인다. 그러나, 발성 방법의 제약으로 인하여 명령어 인식과 같이 단순한 형태와 인식 시스템에 주로 이용된다. 이에 반해 연속음성은 문장의 형태로 자연스럽게 발성되므로 단어의 구분이 불분명하여 단어간 조음 현상으로 인한 변이가 심하다. 또한, 구분 발성과 달리 같은 단어라도 문장 내에서의 발성 속도의 차나 단어내의 변이도 심하게 된다. 일반적으로 연속음성은 고립단어에 비해 발성의 변이가 크고 단어간의 조음 결합 현상과 단어 경계의 불분명성 등으로 인하여 인식이 훨씬 어렵다고 알려져 있으나 보다 자연스럽게 음성인식을 사용할 수 있기 위해서는 연속음성의 인식이 필수적이다. 현재 구분발성 단어에 대한 음성인식은 실용화가 가능한 수준에 이르렀으며 이미 여러 제품들이 상품으로 나와 있는 상태이나 연속음성 인식은 아직 제한된 영역에서나 이용되고 있으며 실용화하기에는 많은 문제점들을 가지고 있다. 따라서 앞으로도 음향, 언어, 대화 처리 등 여러 분야에서 기술 개발이 필요하다.

대부분의 대용량 연속음성인식 시스템에서는 기본 단위 모델로 단어보다는 subword를 사용한다. 기본인식 단위로는 주로 음소나 음소와 유사한 단위들을 주로 사용하게 되며, 이러한 음성인식 시스템에서는 단어나 문장

* 한국과학기술원 전산학과
접수일자 : 1996년 11월 7일

을 인식하기 위해서 기본 subword 단위들로 구성된 단어 발음사전(pronunciation lexicon)을 필요로 하게 된다. 단어 발음사전은 일반적으로 단어에 해당하는 표준 발음표기를 인식단위의 열로 나열함으로써 구성하거나 음운학적 지식을 가진 전문가에 의해서 만들어진다. 이렇게 구성된 발음사전은 몇 가지 문제점들을 갖게 된다. 첫째, 표준 발음 표기로 구성된 단어사전은 여러 화자들의 발성의 변이를 표현하기 힘들다는 문제점이 있다. 화자의 발성 변이는 크게 아래와 같이 2가지로 나눌 수 있다.

1. 단어내 변이 : 개인의 발성 습관과 지역에 따른 사투리 등으로 개개의 발성에 차이를 갖게 되며, 문장내의 위치나 내용에 따라 발성 속도의 차가 생긴다. 단어내 변이는 표준 발음 표기만으로는 해결할 수 없기 때문에 일반적으로 복수개의 발음 표기를 이용하여 해결한다.
2. 단어간 변이 : 단어와 단어의 경계에서 조음결합 현상에 의해 발성하는 변이로 주로 단어의 첫 음소와 마지막 음소의 음가가 주로 변하게 된다. 특히 단어간 변이는 구분 발성 단어와 달리 뒤에 오는 단어의 영향을 받아 음가가 변하므로 복수개의 발음 표기만으로는 해결하기가 힘들므로 단어 주위의 문맥에 대한 고려가 필요하다.

발성 변이에 대한 해결 방법으로 [7]에서는 복수개의 발음표기를 사용하고 있다. 복수개의 발음 표기는 수동으로 작성된 후 인식기에 의한 적응 과정을 거친 후 각 발음 표기에 대한 확률을 결정하게 된다. 또 [8]에서는 단어간 변이를 해결하는 방법으로 발음 규칙을 사용하고 있으며, 이 발음 규칙은 표준 발음 표기와 전문가가 음성을 듣고 작성한 발음 표기간의 차를 규칙화한 후 단어간 천이 시 적용시킨다. 그러나 이러한 방법들은 초기에 전문가가 작성한 발음 표기 방법들을 필요로 하기 때문에 발음 사전의 구성이 쉽지 않고, 복수의 발음 표기를 정하는 방법과 표현 방법에 따라 인식율의 향상에 많은 영향을 미친다는 문제점이 있다.

둘째, 인식의 기본 단위로 음소와 같은 언어적(linguistic)인 단위를 사용하지 않고 음소 유사 단위(PLU : phone like unit)나 acoustic unit과 같이 음향학적으로 유사도에 기반한 단위를 쓰는 경우 언어적인 단위나 단어로의 사상이 필요하므로 발음사전을 만들기 위해서는 많은 노력을 필요로 하게 된다. 일반적으로 음향학적인 단위를 이용하여 발음 사전을 구성하기 위해서는 인식기에 의해 각 단어에 대한 최적의 단위열을 구한 후 발음 사전을 구성하고 이 발음 사전을 이용하여 다시 인식기에 사용되는 단위 모델을 학습을 반복하는 방법을 사용하게 된다. 이밖에 음향학적 모델과 음소간의 사상관계를 구한 후 표준 음소에 대한 발음 표기를 음향학적 단위에 대한 발음 표기로 바꾸는 방법을 사용하기도 한다.

셋째, 일반적으로 음성인식기에서 사용되는 기본 subword 단위들은 모델링 특성상 발음사전에 사용되는 단위들과 정확히 일치하지 않기 때문에 연속음성인식기의 성능을 최적화 할 수 있는 발음사전의 구성이 어렵게 된다. 이는 발음사전에서는 언어적으로 정확히 구분되는 자소(grapheme) 단위를 사용하는데 비해 음성인식기에는 음소(phone)와 같은 언어적 단위와는 다소 차이가 있는 음향학적으로 학습된 단위 모델들을 사용하는데 원인이 있다. 따라서 연속음성인식기의 성능 향상을 위해서는 발음사전과 음성인식기가 대상으로 하는 단위 모델 사이의 불일치를 해결할 방법이 필요하다. [4]에서는 음성인식기의 음소열 인식 결과를 이용하여 발음 사전을 구성하는 방법을 제안하였다. 여기서 각 단어의 발음 표기는 인식기로부터 탐색과정을 거쳐 얻은 다수의 최적 음소열을 군집화 하여 하나나 복수개의 표기를 얻고 있다. 그러나 이 방법은 인식기에 의존하여 각 음성에 해당하는 최적의 음소 표기열들을 구하기는 하지만 이를 인식에 사용하기 위해서 효과적으로 병합하거나 고르는 과정이 요구된다.

이상에서 설명한 발음사전의 문제점을 극복하기 위해서 본 논문에서는 확률 발음사전을 제안하였다. 확률 발음사전은 학습 음성자료로부터 인식 모델과 인식기의 특성을 반영하여 학습되며, 각 단어의 발음 표기 모델은 기본 단위 모델과 같은 HMM(hidden Markov model)으로 표현된다. 제안한 방법은 단어의 발성 변이들을 효과적으로 수용할 수 있으며, 음향학적 단위를 이용한 발음 표기에도 적용이 가능한 모델이다. 특히 발음사전이 음성인식기와 단위모델의 특성에 맞게 구성됨으로써, 실험결과 표준 발성 표기법을 사용한 연속음성인식기에 비해 높은 인식율 향상을 얻을 수 있었다.

논문의 구성은 다음과 같다. 2장에서는 개발된 연속음성인식 시스템의 기본적인 구성에 대해서 기술하고 3장에서는 확률 발음사전의 구성과 인식시스템에 적용에 대해서 기술한다. 4장에서는 인식 시스템의 성능 평가 및 실험 결과를 분석하고 5장에서 결론을 맺는다.

Ⅱ. 연속음성인식 시스템

이 장에서는 본 연구에서 사용한 기본 연속음성인식 시스템에 대해 살펴보고자 한다. 시스템은 그림 1과 같이 구성되어 있으며, 이 중에서 핵심 요소인 단위 모델, 발음사전, 언어 모델 구성 방법과 탐색 알고리즘에 대해 기술하고자 한다.

2.1 단위 모델

음성인식은 음향학적인 신호를 언어적으로 해독하는 과정으로, 인식에 사용되는 인식 단위에 따라 인식 시스템의 구성 및 인식 방법에 많은 차이를 갖게 된다. 널리 사용되는 인식 단위로는 음소, 음절, 단어 등의 언어적으로 정의된 단위들과 음소 유사 단위(PLU : phone like unit)

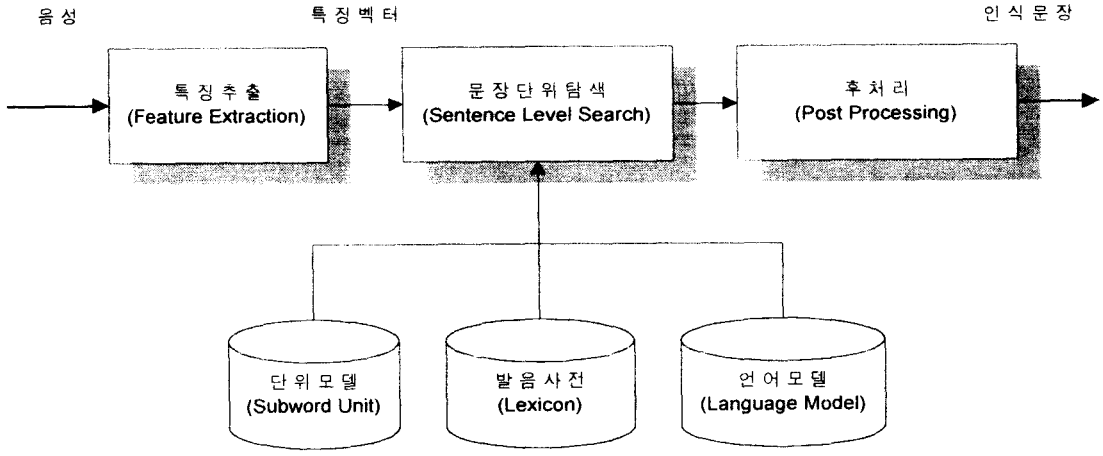

그림 1. 연속음성인식 시스템의 구성

나 음향학적인 유사도에 기반한 단위들이 사용되고 있다. 언어적으로 정의된 단위는 인식 대상 어휘를 위한 발음 사전(lexicon) 구성이 용이한 장점이 있으나, 발성음으로부터 해당 단위를 수동이나 자동으로 분할을 통해 수집하기가 어려운 점이 있다.

특히 대어휘 연속 음성인식기에서 기본 인식 단위로 주로 사용되는 음소(phone)는 단어나 음절에 비해 그 종류가 작고 학습에 필요한 충분한 자료를 모으기가 용이하다는 장점이 있다. 그러나 음소는 좌우에 위치하는 음소에 영향을 많이 받으므로, 이를 고려하기 위해서 세분화된 문맥 의존(context dependent) 음소 모델을 구성하기도 한다. 문맥 독립(context independent) 음소는 문맥 의존 음소에 비해 많은 변이를 포함하므로 모델링이 어려워지고 인식율에 있어서도 저조한 결과를 보인다. 따라서 문맥 독립 음소를 사용할 경우 단위 모델에 대한 정확한 모델링 뿐만 아니라 분별 학습, 후처리 등의 충분한 뒷받침 없이는 높은 인식율을 기대하기가 어렵다. 본 논문에서는 기본 단위 모델로 표 1의 36개의 문맥 독립 음소와 이를 바탕으로 한 3017개의 문맥 의존 음소 모델인 triphone을 사용하여 비교 실험하였다.

음소 모델들은 [3]에서 제안한 HMVQM(hidden Markov VQ model)을 사용하여 모델링 하였다. HMVQM은 HMM과 마찬가지로 정해진 상태 수와 입력 파라미터를 이용하여 한 상태에서 다른 상태로 천이 하는 상태 천이 프로세스와 한 상태 안에서 입력 파라미터를 관측할 관측 프로세스의 2가지 랜덤 프로세스로 시간에 따른 음성 패턴의 변화를 모델링 하게 된다. 이러한 HMVQM은 식(1)과 같이 정의된다. HMVQM은 각 상태마다 작은 크기의 양자화 코드북을 두고 출력확률로 입력 파라미터와의 양자화 왜곡거리를 사용하므로 연속(continuous)분포 HMM의 특수한 형태로 볼 수 있다. HMVQM의 장점은 연속형 HMM에 비해 파라미터의 수가 적기 때문에 적은 음성자료로도 학습이 가능하며 고속 코드북 탐색 방법을 사용함으로써 계산량을 줄일 수 있고, 이산(discrete) 분포 HMM에 비해서는 모델들 간의 분별력을 높일 수 있다. 그림 2는 음소 모델에 사용되는 3개의 상태를 갖는 left-to-right HMVQM 모델로, 상태간의 jump는 허용하지 않는다.

$$\lambda = \{A, Q, N\}$$

$$A = state\ transition\ probability \quad (1)$$
$$Q = state\ VQ\ codebook$$
$$N = number\ of\ states$$

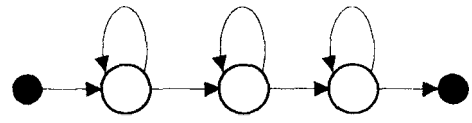

그림 2. 음소 모델

표 1. 음소 기호의 정의

자음	ㅂ	b	ㄷ	d	ㄱ	g	ㅈ	j
	ㅃ	P	ㄸ	T	ㄲ	K	ㅉ	C
	ㅍ	p	ㅌ	t	ㅋ	k	ㅊ	c
	ㅅ	s	ㅆ	S	ㅎ	h		
	ㄹ	r	ㄴ	n	ㅁ	m	ㅇ	N
모음	ㅏ	a	ㅑ	ya	ㅘ	wa		
	ㅓ	v	ㅕ	yu	ㅝ	wv		
	ㅗ	o	ㅛ	yo				
	ㅜ	u	ㅠ	yu				
	ㅡ	U			ㅢ	Wi		
	ㅣ	i			ㅟ	wi		
	ㅔ, ㅐ	e	ㅖ	ye	ㅞ, ㅙ, ㅚ	we		

2.2 발음 사전

음소를 기본 인식 단위를 사용하는 연속음성인식기의 개발에서는 발음사전의 구현이 필수적이라 할 수 있다. 한글의 자소(grapheme)는 구체적인 음운 현상을 반영한 음소단위의 실현(phone)이 아니므로 단어는 음소, 문맥의 존형 음소, 변이음 등의 음성인식 단위를 정의한 후 한글 읽기 규칙에 따라 소리나는 대로 정의된 발음 기호로 바꾸게 된다. 단어의 자소를 음소로 바꾸기 위해서는 변환 규칙과 예외 발음 사전이 필요하며, 하나의 표준 발음 표기만으로는 단어의 발음을 표기 할 수 없는 경우 여러 개의 발음 표기를 두기도 한다.

기본 연속음성인식 시스템에서는 표 2에서 보듯이 각 단어는 변환규칙과 예외 발음 사전을 이용하여 자동으로 음소의 열을 생성하였으며, 각 단어마다 기본적으로 하나의 표준 발음 표기만을 사용하였다.

표 2. 단어의 표준 발음 표기 예

단 어	표준 발음 표기
가	/ g a /
가격	/ g a g yv g /
가격대	/ g a g yv g T e /
가격상승	/ g a g yv g S a N s U N /
가격인상	/ g a g yv g i n s a N /
가격인하	/ g a g yv g i n h a /
가격하락	/ g a g yv k a r a g /
가구	/ g a g u /
가까운	/ g a K a u n /
가깝지	/ g a K a b C i /

2.3 언어 모델

대어휘 연속음성인식 시스템에서 언어 모델은 이웃하는 단어사이의 연관성을 나타내는 정보이다. 이 언어 모델은 연속음성을 인식할 때 불필요한 단어나 어절을 제한함으로써 문장의 탐색 시간과 인식율을 높이는 역할을 하게 된다. 음성인식에 주로 사용되는 언어 모델로는 구구조(phrase structure) 문법에 기반한 언어 모델과 통계적 언어 모델을 들 수 있다. 구구조 문법의 경우 정규 문법(regular grammar)이나 문맥 자유 문법(context free grammar)을 사용하여 문장의 탐색과 동시에 parsing을 수행하여 문법의 구조에 어긋난 탐색 공간(search space)을 제거하게 된다. 특히 날짜, 간단한 명령어 등과 같이 비교적 단순한 문법의 경우 FSN(finite state network)을 사용하여 쉽게 표현이 가능하나 단어의 수가 늘어남에 따라 network의 state 수가 급격히 증가하게 되어 탐색 시간이 오래 걸리는 단점이 있다. 이밖에 구구조 문법을 음성인식에 사용할 경우 대상으로 하는 영역의 문장을 표현하는 문법을 작성하기가 쉽지 않고 정해진 문법에 조금이라도 벗어난 문장은 인식이 어렵다는 단점이 있다.

이에 비해 통계적 언어 모델은 일반적으로 주어진 영역의 많은 텍스트 분상으로부터 쉽게 추출이 가능하고, 입력 문장 전체를 parsing하지 않고 문장의 발생 확률만을 계산하므로 학습 문장과 부분적으로 다른 문장도 인식할 수 있는 장점이 있다. 통계적 언어 모델에서는 임의의 단어열 W가 가지는 확률 P(W)를 계산한다. 그러나 식(2)의 확률은 실제로 계산하기가 거의 불가능하므로 N-gram 단어 모델을 사용한다. 이 모델은 식(3)과 같이 P(W)를 이전에 발생한 N−1개의 단어만을 이용하여 계산하는 것이며, 통상적으로 N은 2(bigram) 또는 3(trigram)을 많이 사용한다. 통계적 언어 모델은 단어간의 확률이 아니라 단어의 종류(word class)에 따라 만듦에 따라 문법적 제약에 필요한 파라미터의 수를 줄이고 학습 문장 자료를 효율적으로 활용할 수 도 있다.

$$P(W) = P(w_1, w_2, ..., w_L) = P(w_1) \prod_{i=2}^{L} P(w_i | w_1, ..., w_{i-1}) \quad (2)$$

$$P_N(W) = P(w_1, ..., w_{N-1}) \prod_{i=N}^{L} P(w_i | w_{i-N+1}, ..., w_{i-1}) \quad (3)$$

본 논문에서는 기본 인식 시스템의 언어 모델로 단어의 bigram 문법을 사용하였다. 이때 N개의 단어에 대한 bigram을 행렬(matrix)로 구현할 경우 필요한 메모리는 $O(N^2)$가 필요하게 되므로 대규모의 어휘 인식 시스템에서는 단어 bigram을 이용하기가 쉽지 않다. 그러나 일반적인 문장의 단어 복잡도(perplexity)는 N보다 매우 작으므로 단어 bigram을 희소 행렬(sparse matrix)로 구현하면 적은 메모리로도 이용할 수 있게 된다.

2.4 탐색 알고리즘

연속음성 인식은 그림 1의 단위 음성 모델, 발음 사전, 언어 모델 등의 지식 정보를 이용하여 식(4)을 최적화 하는 단어열을 찾는 과정이다. 여기서 식(4)의 $P(X|W)$는 단위 음성 모델과 발음사전에 의한 음향학적 확률이고, $P(W)$는 언어모델에 의한 단어열의 확률이다.

이러한 최적의 단어열을 찾는 과정은 지식정보로 구성된 상태 공간(state space)을 탐색하는 알고리즘을 사용하여 수행된다. 여기서 상태 공간은 그림 3과 같이 각 단어에 포함된 음소 노드들로 구성되며 다시 단위 음성 모델의 상태들로 이루어진 network으로 표현된다. 탐색을 통해 network으로부터 단위 음성 모델 상태들의 최적 경로를 추적하게 되고 최적 상태 경로를 이용하여 음소, 단어, 문장 순으로 결과를 만들어 간다.

$$W^* = \underset{W}{\operatorname{argmax}} \; P(X|W)$$

$$= \underset{W}{\operatorname{argmax}} \; P(X|W)P(W) \quad (4)$$

본 논문에서 사용되는 기본적인 탐색 알고리즘은 1-pass DP(dynamic programming)[1]에 기반한 tree-trellis 탐색 알

고리즘[6]을 사용한다. tree-trellis 탐색 알고리즘은 N개의 최적 후보 문장을 찾는데 있어서 1개의 최적 문장을 찾는 데 비해 거의 메모리와 시간의 증가 없이 효율적으로 찾을 수 있는 탐색 방법중 하나이다. tree-trellis 탐색 알고리즘의 구성은 time synchronous trellis 탐색에 해당하는 전방향 탐색과 time asynchronous tree 탐색에 해당하는 역방향 탐색으로 이루어 된다. 먼저 전방향 탐색에서는 1-pass DP 알고리즘을 이용하여 부분 경로(partial path)에 대한 확률 값을 저장한 후에 역방향 탐색에서 A* 알고리즘을 이용하여 최적의 부분 경로들부터 확장해 가면서 N개의 최적 문장을 찾게 된다. 이때 A* 알고리즘에서 사용되는 부분 경로들의 평가 함수(evaluation function) 값은 1-pass DP에서 구해진 전방향 부분 경로 확률 값과 현재까지의 역방향 부분 경로 확률 값의 합을 사용하여 부분 경로 상에서 전 경로(full path)에 대한 정확한 확률 값을 알 수 있게 되므로 불필요한 부분 경로의 확장이 줄게 된다. 따라서 tree-trellis 탐색은 일반적인 stack 알고리즘에 비해 메모리의 증가가 적으면서도 빠르게 최적의 문장들을 찾을 수 있게 된다.

본 논문에서는 전방향 탐색 시 전체 탐색 공간에 대한 고려로 생기는 계산량의 비효율성을 개선하기 위해 beam 탐색 기법을 사용하였다. beam 탐색에서는 매 입력 프레임마다 모든 후보의 경로들을 확장하지 않고 확률이 높은 일부 후보 경로들만을 확장하게 된다. 이때 beam 탐색의 임계값을 작게 하면 탐색의 정확도는 감소하나 계산량이 줄어들게 되며, 크게 하면 그 반대가 된다. 따라서 beam 탐색에서는 정확도를 유지하면서 계산량을 줄이기 위해서는 임계값을 적절하게 선택해서 매 시간 마다 탐색 공간을 적당하게 유지시켜야 한다. 일정한 탐색 공간을 유지시키기 위한 방법으로는 현재 상태들의 확률 값을 고려하여 임계값을 조정하는 방법과 활성 상태의 수를 현재의 확률 값에 따라 제한 방법이 있으나 본 논문에서는 활성 상태 수를 제한하는 방법을 사용하였다.

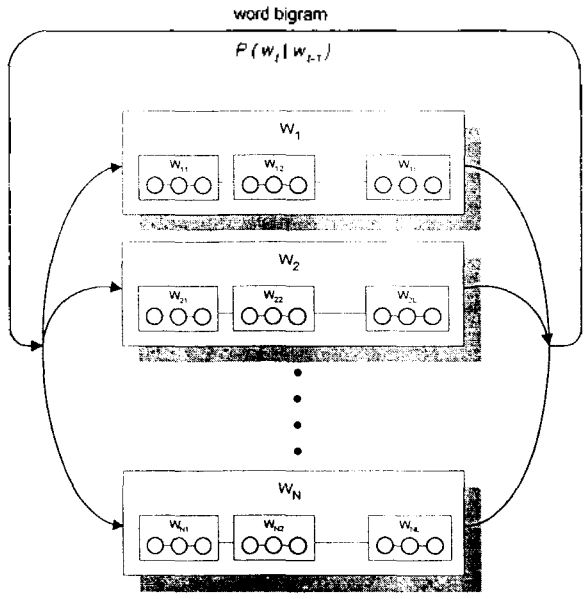

그림 3. 단어 network

Ⅲ. 확률 발음사전 모델

표준 발음 표기만으로는 실제 음성 자료를 정확히 표현하지 못하기 때문에 각 단어의 발음 변이에 대한 고찰이 있어야 한다. 그러나, 전문가 지식이나 음운 규칙을 기반으로 하는 방법은 많은 노력을 필요로 할뿐 아니라 실제 음성자료에서 나타나는 발성의 변이를 모두 표현하기가 힘들다. 또한 발성의 변이에 대한 표현도 인식기에서 사용되는 음소모델 보다는 언어적인 음소단위에 대한 표현이기 때문에 인식 시스템의 최적화와는 거리가 멀게 된다. 본 논문에서 제안하고 있는 확률 발음사전 모델은 각 단어의 발음에 대한 변이를 직접 음성 자료와 음소모델로부터 자동으로 구하게 된다.

그림 4(b)에서 보듯이 확률 발음사전은 HMM과 같이 모델링 된다. HMM은 단위음소 상태들의 Markov chain으로 이루어져 있으며 각각의 단위음소 상태는 음소들에 대한 확률분포 함수를 포함하게 된다. 확률 발음사전은 음성 변이에 대한 모델링 관점에서 복수개의 발음 표기법을 사용하거나 그래프 형태로 구성된 발음사전을 사용하는 경우와 유사하지만 기존의 발음 표기는 인식을 위해 그림 4(a)와 같이 각각의 음소들이 Markov state들로 확장되어 사용되는 반면 확률 발음사전에서는 그림 4(b)와 같이 음소의 발생이 확률적으로 결정되므로 hidden Markov state들로 표현된다. 즉, 기존의 발음사전은 음소의 network으로 단어를 표현하고 있기 때문에 결정적(deterministic)인 반면 제안된 확률 발음사전은 단어를 확률적(stochastic)인 음소 상태(phone state)로 모델링하고 있다는 점에서 차이가 있다.

이 장에서는 연속음성인식에 사용되는 확률 발음사전을 구성하는 방법과 인식과정에서 확률 발음사전을 이용하는 방법에 대해서 기술하고자 한다.

3.1 확률 발음사전의 생성

확률 발음사전은 HMM으로 표현되기 때문에 이를 구성하기 위해서는 단위음소 상태들의 Markov chain과 각 상태에서의 확률분포 함수를 추정해야 한다. 먼저 HMM의 상태 수와 형태는 표준 발음 표기를 기반으로 하여 표준 발음 표기에서 나타난 음소의 수를 상태의 수로 정하며 이 음소들이 left-to-right 형태로 연결된 형태로 구성하게 된다. 또, 각 상태에서의 확률분포 함수는 음성인식기에 의해 음소단위로 자동으로 분할된 각각의 음성 자료들과 음소 모델을 이용하여 유사도(likelihood)를 구한 후 확률 분포를 추정하게 된다.

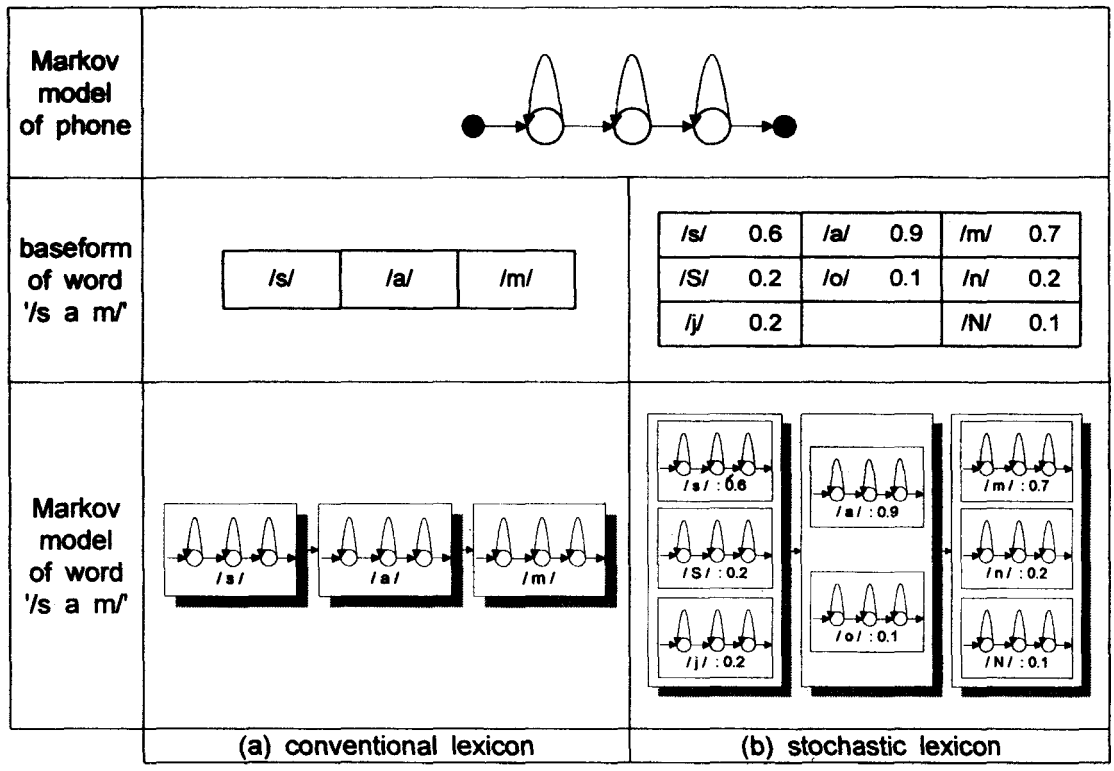

그림 4. 기본 발음사전과 확률 발음사전의 비교

음성의 자동 분할과 음소모델 학습을 위해서는 segmental K-means 학습방법을 사용하였다[5]. segmental K-means는 주어진 음소 모델, 음성자료와 단어에 대한 음소열로부터 음성을 최적의 음소 분절(segment)들로 분할된다. 이렇게 분할된 분절들은 다시 음소 모델을 학습하는데 이용되며, 각 단어의 음소 상태 별로 모아진 분절들과 학습된 음소 모델을 이용하여 각 상태별로 음소의 확률 분포를 학습하게 된다. 따라서 발음사전은 segmental K-means 학습방법을 이용한 음소 분할과 음소모델 학습, 음소변이에 대한 확률 분포 추정에 의해 자동으로 생성된다.

segmental K-means 학습방법과 확률 발음사전에서 음소상태의 확률 분포 추정에 대한 방법은 아래와 같다.

segmental K-means 알고리즘을 이용한 학습 방법

1. 초기화(Initialization)
   초기모델(bootstrap model)을 작성한다. 초기모델은 HMM의 상태 수에 따라 문장을 평균분할 하거나, 수동 분할된 음소자료를 이용한다.
2. 분할(Segmentation)
   단어의 음소 표기열로부터 문장을 network으로 구성한 후, Viterbi 정합을 이용하여 해당 음성자료와 주어진 HMM간의 likelihood를 최적화 하는 음소열로 분할한다.
3. 추정(Estimation)
   음소별로 분할된 음성자료로부터 각 음소 모델의 파라미터를 추정한다.
4. 반복(Iteration)
   2와 3 과정을 모델 파라미터가 수렴할 때까지 반복한다.

확률 발음사전에서 음소 상태의 확률분포 추정 방법

1. 기본단위 분할
   segmental k-means 학습 방법을 이용하여 문장을 최적의 음소 단위열로 분할한 후 각각의 음소 모델들을 학습한다.
   음성자료 S_{ij}은 i번째 단어의 j번째 음소에 대한 분할된 자료의 집합이다.
   $S_{ij}=\{S_{ij}^1, S_{ij}^2, ..., S_{ij}^K\}$, K; 전체 분할 자료의 개수
2. 음소단위 인식
   i번째 단어의 j번째 음소 노드 W_{ij}에 대해서 분할된 각각의 음성자료 S_{ij}^k를 이용해서 음소인식을 수행하고, 각 음소 모델에 대한 퍼지 유사도 P_{ijv}^k를 구한다.

$$P_{ijv}^k = p(\lambda_v | W_{ij}, S_{ij}^k)$$

$$\approx \left[\sum_{p=1}^{P} \left[P(S_{ij}^k|\lambda_v)/P(S_{ij}^k|\lambda_p) \right] \right]^{-1} \quad (5)$$

$$\sum P_{ijv}^k = 1$$

$$P(S_{ij}^k|\lambda_v) = -\log p(S_{ij}^k|\lambda_v) \quad (6)$$

where, $L=\{W_1, W_2, ..., W_N\}$; 어휘 사전,
N; 전체 단어 개수
$W_i=\{W_{i1}, W_{i2}, ..., W_{iM}\}$; 단어의 음소 상태열,
M; 단어의 음소 수
$\lambda=\{\lambda_1, \lambda_2, ..., \lambda_P\}$; 음소 모델,
P; 전체 음소 개수

3. 확률 발음 사전 구성

각 음소 상태 W_{ij}에 대한 퍼지 유사도 P_{ijv}^k로 부터 출력 확률 분포를 구한다.

$$P_{ijv} = p(\lambda_v|W_{ij}) = \frac{1}{K}\sum_{k=1}^{K} P_{ijv}^k$$

$$\sum_v P_{ijv} = 1 \quad (7)$$

3.2 확률 발음사전을 이용한 인식

앞에서 설명한 것과 같이 연속음성을 인식하기 위해서 사용되는 전방향 탐색 방법은 beam 탐색 기법을 이용한 1-pass DP 알고리즘이다. 이때 탐색 공간은 그림 3과 같이 network으로 표현되며, 단어의 수가 늘어남에 따라 network을 구성하는 음소의 노드 수와 HMM의 상태의 수가 증가하게 된다. 만약 복수개의 발음 사전을 사용하는 경우라면 음소 노드가 늘어나게 되므로 network의 상태 수는 더욱 커지게 되며, 이는 인식의 속도와 정확성에 중요한 영향을 미치게 된다. 그러나 확률 발음사전에서는 표준 발음사전을 사용하는 경우와 같은 수의 상태를 갖게 되므로 탐색 공간의 증가는 없게 된다.

탐색 과정에서는 상태의 경로를 알 수 없기 때문에 매 시간마다 단어의 경계들에 대한 확률 값을 구해야 한다. 각 상태에서의 확률을 구하기 위해서는 Viterbi 알고리즘을 이용하여 식(8)과 같이 음소 노드 W_{ij}에 대한 음성 파라미터 벡터 $X_1, ..., X_t$의 확률을 구한다.

$$p(X_1, ..., X_t|W_{ij}) = Q_{ijS(W_{ij})}(t)$$

$$Q_{ijk}(t) = \max_s [a(s_k|s')D_{ijk}(X_t)Q_{ijk}(t-1)] \quad (8)$$

여기서 인자 i는 단어를 j는 음소 노드를 k는 음소 HMM의 상태를 나타내며, $S(W_{ij})$는 각 단어에 마지막에 해당되는 HMM의 상태를 $a(s|s')$는 HMM의 상태 천이 확률을 의미한다. $D(X)$는 각 HMM 상태에서의 likelihood를 나타내며, 만약 단일 발음 표기를 사용하는 경우 식(9)과 같이 각 음소 노드에 해당하는 HMM의 상태에서만 likelihood를 계산하게 된다. 그러나 확률 발음 사전을 사용하는 경우는 식(10)과 같이 음소 노드에서의 음소 유사도 P_{ijv}와 각 음소모델의 likelihood 값의 곱의 합으로 계산된다. 이때 식(10)은 식(9)을 사용할 때보다도 likelihood 값을 계산하는데 더 많은 계산을 필요로 하게 된다. 특히 이산 분포 HMM보다는 연속 분포 HMM에서는 이 부분의 계산량이 차지하는 비중이 크기 때문에 모든 음소에 대해 likelihood를 계산하지 않고, 각 음소 노드에서 유사도 P_{ijv}가 높은 N개의 음소에 대해서만 계산을 하는 방법으로 연산량을 줄일 수 있다.

$$D_{ijk}(X_t) = p(X_t|W_{ij}, s_k) = \begin{cases} p(X_t|s_k, \lambda_v) & \text{if subword model of } W_{ij} \text{ is } \lambda_v \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

$$D_{ijk}(X_t) = p(X_t|W_{ij}, s_k) = \sum_{v=1}^{P} P_{ijv} \cdot p(X_t|s_k, \lambda_v) \quad (10)$$

Ⅳ. 실험 및 결과

3.1 실험 환경

본 논문에서 제안한 방법의 성능평가를 위해서 사용된 음성 데이터 베이스는 한국과학기술원 통신 연구실에서 제작한 무역상담용 연속음성 데이터 베이스[9]이다. 표 3의 예와 같이 문장은 품사에 의해 분류된 3016개의 어휘로 구성되어 있으며, 남성 100명, 여성 50명이 평균 98 문장씩을 자연스럽게 발성하였다. 이중 음소모델과 발음사전의 학습을 위해서 남성 75명, 여성 25명이 발성한 문장을 사용하였으며, 평가를 위해서 나머지 남성 25명, 여성 25명이 발성한 문장을 사용하였다. 실험에 사용한 특징 파라미터는 14차 멜캡스트럼 계수와 14차 델타 멜캡스트럼 계수, 에너지, 델타 에너지를 함께 사용하였다.

표 3. 연속음성 데이터 베이스에 사용된 문장의 예
(문장에 사용된 어휘별로 띄어쓰기)

독일어 로 된 간단한 팜프렛 이 두 개 있습니다
이것 은 그것 과 가장 비슷한 제품 은 아니지만 품질 은 더 훌륭합니다
그 생산공장 혼자서는 그 일 을 다 감당 해 낼 수 가 없습니다
사십 파운드 가 최저가격 입니까
무엇 을 도와 드릴까요
선적품 의 품질 이 최초 의 제품모델 보다 좋지 못했습니다
다스캣 의 이 차 선적 을 연기 하셔야 할 것 같습니다
대략 언제쯤 될지 알 수 없습니까
비직항선 인 경우 화물 은 홍콩 에서 환적 이 됩니까
그 선사 에는 방콕 으로 가는 배 가 자주 있을까요
그렇게 지연 되는 이유 가 뭐죠

인식에 사용된 기본 단위는 36개의 문맥 독립 음소와 3017개의 문맥 의존 음소(triphone), 2개의 무음을 사용하였다. 음소를 모델링 하기 위해 사용한 방식은 left-to-right 형태의 HMVQM(hidden Markov VQ model)이다. 음소단위로 HMVQM을 구성하였으며, 각 음소모델은 3개의 state로 구성되며 문장의 처음과 끝, 그리고 단어사이의 무음은 1개의 state로 구성된다. 각 sate의 출력 확률분포를 추정하기 위해서 표 4와 같은 codebook을 사용하였다.

표 4. 음소모델의 구성

		모델 수	state 수	codeword 수
문맥 독립 음소 (phone)	음소 모델	36	3	18
	무음 모델	2	1	10
문맥 의존 음소 (triphone)	음소 모델	3017	3	10
	무음 모델	2	1	8

3.2 실험 결과 및 검토

연속음성의 인식율은 단어의 인식율과 문장 인식율로 나타내었다. 단어의 오류는 치환, 첨가, 삭제에 의한 오류를 포함하여 식(11)과 같이 구하며, 문장 인식율은 단어오류가 포함되지 않은 인식 결과만을 이용하여 구하게 된다.

$$Word\ Accuracy = \frac{Correct - Ins}{Correct + Subs + Dels} \times 100\ (\%) \qquad (11)$$

표 5는 기존의 발음사전을 사용했을 경우와 제안된 확률 발음사전을 사용했을 경우를 비교 실험한 결과를 보여주고 있다. 먼저 문맥 독립 음소인 phone 모델을 사용했을 때 보다 문맥 의존 음소인 triphone 모델을 사용할 경우 단어의 오류는 72.3%를 문장의 오류는 46.0%를 줄일 수 있었다. 발음사전에 따른 인식결과는 제안된 확률 발음 사전을 사용했을 경우 단어의 오류는 23.6(phone), 22.4%(triphone)를 문장의 오류는 6.8(phone), 10.0%(triphone)를 줄일 수 있었다. 이 결과에서 보듯이 연속음성 인식에서 인식단위의 모델링 못지 않게 발음사전의 구성이 인식 성능에 중요한 영향을 미침을 알 수 있다. 특히 제안된 발음 사전은 조음 결합 특성을 어느 정도 수용할 수 있는 triphone과 같은 문맥 의존 음소에서도 비슷한 성능 향상을 보임을 알 수 있다.

표 5. 연속음성 인식 결과

lexicon	subword unit	word accuracy(%)	sentence accuracy(%)
conventional lexicon	phone	58.13	26
	triphone	88.40	60
stochastic lexicon	phone	68.00	31
	triphone	91.11	64

표 6은 확률 발음 사전을 이용할 경우 계산량의 증가를 가져오게 되므로 이를 개선시키기 위해서 인식 시 확률 발음 사전 중 발생 확률이 높은 상위 N개의 음소만을 이용하여 실험한 결과이다. 실험 결과 모든 음소를 고려하는 경우보다는 N을 5 정도로 사용하는 경우가 인식율이나 인식 시간에서 유리하였다. 이는 발음의 변이가 그림 6과 같이 몇몇 음소들로 결정되고, 그 수가 전체 음소 수에 비해 크지가 않기 때문으로 판단된다. 또, 그림 5(c)에

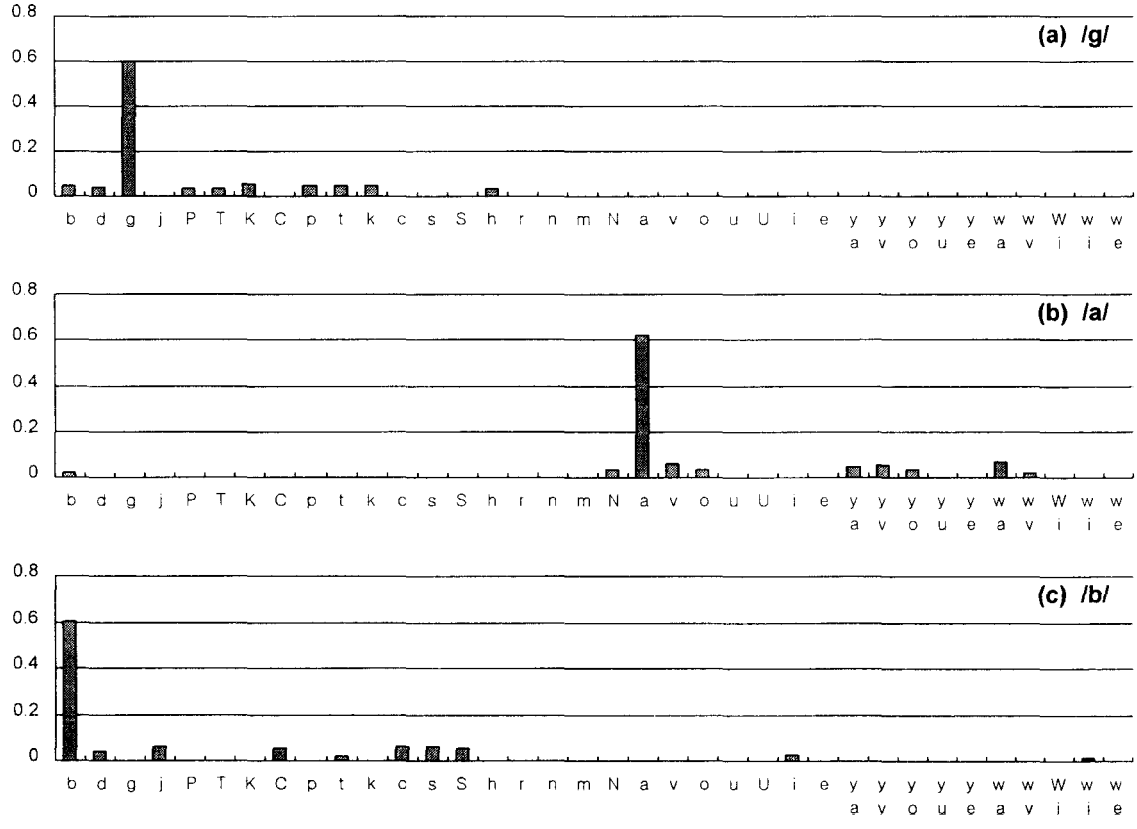

그림 5. 학습된 확률 발음 표기의 예 (값 : /g a b/)

표 6. 확률발음사전에 사용된 음소 수에 따른 성능 평가(phone unit 사용)

N-best unit	word accuracy (%)	sentence accuracy (%)	recognition time (sec)
N=1 (conventional)	58.13	26	7.55
N=3	66.42	31	17.51
N=5	68.00	31	18.95
N=10	67.20	30	23.45

서 보듯이 단어 '값'의 마지막 음소를 표준 발음 표기를 사용할 경우 'ㅂ' 발음으로밖에 표현할 수 없으나, 확률 발음 표기에서는 뒤에 모음으로 시작되는 단어들이 올 때는 발생하는 'ㅂ'과 'ㅅ'의 음가가 모두 포함할 수 있음을 알 수 있다.

V. 결 론

본 논문에서는 대어휘 연속음성인식을 위한 발음사전의 구성에 대해서 기술하였다. 제안된 확률 발음 사전은 단어내 변이와 단어간 변이를 모두 효과적으로 표현할 수 있었으며, 인식 모델과 인식기의 특성을 반영함으로써 전체 인식 시스템의 성능을 보다 높일 수 있었다. 실험 결과 확률 발음 사전을 사용함으로써 단어 오류율은 23.6%, 문장 오류율은 10% 까지를 감소시킬 수 있었으며, 문맥독립 음소뿐만 아니라 문맥의존 음소 모델과 같이 조음결합 특성을 많이 포함하는 단위에서도 비슷한 성능 향상을 얻을 수 있었다. 그러나 제안한 발음사전 모델은 인식 시 기존의 방법에 비해 계산량이 증가되는 단점이 있다. 앞으로 시스템의 인식 시간을 개선하여 실시간 구현을 위해서 불필요한 음소에 관한 계산을 줄일 수 있는 탐색 알고리즘 등에 관하여 계속 연구해 나아갈 것이다.

참 고 문 헌

1. H. Ney, D. Mergel, A. Noll, A.Paeseler, "Data Driven Organization of the Dynamic Programming Beam Search for Continuous Speech Recognition", IEEE Trans. on Signal Processing, Vol.40, No.2, pp. 272-281, Feb. 1992.
2. L.R.Bahl, P.F.Brown, P.V.de Souza, R.L. Mercer, M.A. Picheny, "A method for the Construction of Acoustic Markov Models for Words", IEEE Tans. Speech and Audio Processing, Vol.1, No.4, pp. 443-452 Oct. 1993.
3. Seong Jin Yun and Yung Hwan Oh, "Performance Improvement of Speaker Recognition System for Small Training Data", Proc. ICSLP'94, pp. 1863-1866, Yokohama, 1994.
4. Torbjfrn Svendsen, Frank K. Soong, Heiko Purnhagen, "Optimizing Baseforms for HMM-Based Speech Recognition", Proc. Eurospeech'95, pp. 783-785, Madrid, 1995.
5. Lawrence Rabiner, Biing-Hwang Juang, Fundamentals of Speech Recognition, Prentice-Hall, 1993.
6. Frank K Soong, Eng-Fong Huang, "A Tree-Trellis Based Fast Search for Finding the N Best Sentence Hypotheses in Continuous Speech Recognition", Proc. ICASSP'91, pp. 705-708, 1991
7. Chuck Wooters, Andreas Stolcke, "Multiple-Pronunciation Lexical Modeling in a Speaker Independent Speech Understanding Syatem", Proc. ICSLP'94, pp. 1364-1366, Yokohama, 1994.
8. Nick Cremelie, Jean-Pierre Martens, "On the Use of Pronunciation Rules for Improved Word Recognition", Proc. Eurospeech'95, pp. 1747-1750, Madrid, 1995.
9. 최인정, 권오욱, 박종렬, 박용규, 김도영, 정호영, 은종관, "대용량 한국어 연속음성인식 시스템 개발", 한국음향학회지, 14권, 5호, pp. 44-50, 1995.

▲윤 성 진(Seong Jin Yun)

1992년 2월 : 한국과학기술원 전산학과(학사)
1994년 2월 : 한국과학기술원 전산학과(석사)
1994년 3월~현재 : 한국과학기술원 전산학과 박사과정 재학중
※주관심분야 : 음성인식, 화자인식, 확률모델

▲최 환 진(Hwan Jin Choi)

1990년 2월 : 고려대학교 전산학과(학사)
1992년 2월 : 한국과학기술원 전산학과(석사)
1992년 3월~현재 : 한국과학기술원 전산학과 박사과정 재학중
※주관심분야 : 음성인식, 대화관리, 신경회로망

▲오 영 환(Yung Hwan Oh)

1972년 2월 : 서울대학교 공과대학 전자공학과
1974년 2월 : 서울대학교 교육대학원 공업교육학과(석사)
1980년 3월 : Tokyo Institute of Technology 정보공학전공(박사)
1981년 4월~1985년 6월 : 충북대학교 공과대학 전산학과 조교수
1983년 12월~1984년 11월 : University of California(Davis) 연구교수
1985년 7월~현재 : 한국과학기술원 전산학과 교수로 재직중
※주관심분야 : 음성인식, 음성합성, 음성코딩, 화자인식, 대화관리, 신경회로망, 전문가 시스템