HM-Net을 이용한 한국어 유사음소 단위의 재 정의와 평가

임 영 춘, 오 세 진*, 정 호 열, 정 현 열
영남대학교 전자정보공학부
*대구과학대학 디지털정보통신계열

Definition and Evaluation of Korean Phone-Like Units using Hidden Markov Network

Young-Chun Lim, Se-Jin Oh*, Ho-Youl Jung, Hyun-Yeol Chung
Dept. of Information and Communication Eng., Yeungnam University
*Digital Information and Communication Div., Taegu Science College
E-mail: lyc@speech.yu.ac.kr

요 약

최근 음성인식의 인식 단위로서 문맥의존 음향 모델이 널리 사용되고 있다. 이는 음소의 음향학적 특징, 즉 선행 및 후행음소에 의한 중심 음소의 변이음 모델이 문맥독립 모델보다 좀 더 정확하게 모델링될 수 있기 때문이다. 하지만 강건한 문맥의존 음향 모델을 작성하기 위해서는 모델 파라미터의 병합(tying)과 미지의 문맥(unseen context)의 처리를 위한 좀 더 정교한 해결 방법이 필요하다. 따라서 본 논문에서는 이점을 고려하여 음향학적 특징과 언어학적 특징을 결합하여 상태 분할을 수행할 수 있도록 SSS(Successive State Splitting) 알고리즘의 문맥 방향 상태 분할에 음소결정트리를 접목한 HM-Net(Hidden Markov Network) 구조 결정법을 도입하였다. 또한 HM-Net은 연속적인 상태 분할에 의해 한국어에서 많이 발생하는 변이음들을 효과적으로 모델링할 수 있다는 점을 고려하여 본 연구실에서 기존에 사용하던 48 유사음소 단위에서 문맥의존 음향 모델 작성에 불필요한 변이음을 제거하여 39 유사음소 단위를 재 정의하였다. 도입한 방법과 새로 정의한 유사음소 단위의 유효성을 확인하기 위해 고립 단어, 4연속 숫자음, 연속 음성인식에 대해 인식 실험을 수행한 결과, 모든 실험에서 재 정의한 39 유사음소 단위가 문맥종속형 HM-Net 음향모델을 이용한 한국어 음성인식에 효과적임을 확인할 수 있었다. 특히 연속 음성인식 실험의 경우, 기존의 48 유사음소 단위보다 평균 15.08%의 인식률 향상이 있었다.

1. 서 론

1960년대 이후로 널리 연구되고 많이 사용되는 HMM(Hidden Markov Model)은 시간적, 공간적인 특징을 잘 반영하는 이중 통계적 방법으로 음성인식을 포함한 다양한 분야에서 성공적으로 적용되고 있다[2-8]. HMM으로 음소 단위를 모델링할 때 음소를 구성하는 방법에 따라 문맥독립형 음향 모델과 문맥의존형 음향 모델로 나눌 수 있다. 문맥독립 모델은 대부분 n 상태 m 출력의 단순한 구조로 음소를 독립적으로 모델링하기 때문에 이웃하는 음소에 의한 변이음의 정보를 모두 수용하기에는 부족하다[3-7].

반면에 문맥의존 모델은 문맥독립 모델에 비해 음향 모델의 가지 수는 많지만 이웃 음소에 의한 변이음을 고려한 모델로서 강건한 음향 모델을 생성하는 방법으로 많은 연구가 진행되고 있다[4][6][7]. 하지만 문맥의존 음향 모델은 하나의 중심 음소를 기준으로 선행 및 후행음소에 따라서 수천 가지의 서로 다른 음소가 생성되기 때문에 통계적인 수법인 HMM 기반에서 강건한 음향 모델을 작성하기 위해서는 다양한 문맥 요소가 포함되어 있는 충분한 학습 데이터가 요구된다. 이는 부족한 데이터로는 다양한 변이음을 효과적으로 학습할 수 없을 뿐만 아니라 미지의 문맥 요소가 많이 발생하기 때문이다. 비록 미지의 문맥을 고려한 모델이 문맥독립 모델로 대치될 수 있다고 하지만 인식 성능에는 그리 많은 영향을 미치지 못한다[10][11].

한편 HM-Net은 연속적인 상태 분할에 기반한 HMM의 확장 구조로서 강건한 문맥의존 음향 모델을 작성하는 방법으로 널리 알려져 있으나 이를 이용한 한국어에 대한 연구가 미흡한 실정이다. 따라서 본 논문에서는 HM-Net 구조를 이용한 한국어 음성인식에 관해 검토하고자 한다. 이를 위해 HM-Net을 위한 효과적인 문맥의존 음향 모델을 작성하기 위해 SSS(Successive State Splitting) 알고리즘의 문맥 방향 상태 분할에 음소결정트리를 결합한 HM-Net(Hidden Markov Network) 구조 결정법을 도입한다. 특히 HM-Net은 변이음을 효과적으로 모델링하기 때문에 기존에 사용하던 단일 HMM의 변이음들은 HM-Net 음향 모델의 강건함을 저하시킬 수 있다. 그러므로 HM-Net 모델에 효과적인 유사음소 단위를 다시 정의할 필요가 있다.

본 논문에서는 이상에서 언급한 바와 같이 한국어 음성인식을 위해 도입한 HM-Net 구조 결정법과 이를 위해 새로이 유사음소 단위를 정의한 다음 정의된 유사음소 단위의 유효성을 확인하기 위해 고립 단어, 4연속 숫자음, 연속 음성인식을 대상으로 인식 실험을 수행한 후 그 결과에 대해 보고한다.

2. Hidden Markov Network

일반적인 HMM의 구조가 해당 음소마다 상태들이 독립적으로 n 상태로 이루어져 있는 것과 달리 HM-Net은 하나의 상태 네트워크 구조 속에 모든 문맥의존 모델을 포함하게 되고 유사한 특징을 가지는 상태들을 모델들 사이에서 공유하게 된다

[9][10][11]. 일반적인 HMM에서 모델의 구조를 설정할 때 다양한 실험을 통한 경험적인 요소에 의해 구조를 결정하는 것에 비해 HM-Net은 SSS 알고리즘을 이용하여 연속적인 상태 분할을 통해 모델의 구조를 자동적으로 결정하게 된다[9]. 또한 문맥의존 음향 모델을 효과적으로 작성하기 위해 널리 사용되는 음소결정트리(Phonetic Decision Tree; PDT)를 SSS 알고리즘의 문맥 방향 상태 분할에 적용할 경우 음향학적 특징뿐만 아니라 언어적 특징이 접목되어 모델의 상태 분할이 효과적으로 수행될 수 있다[11]. 이하, 이에 대해 간략히 기술한다.

2.1 PDT-SSS 알고리즘

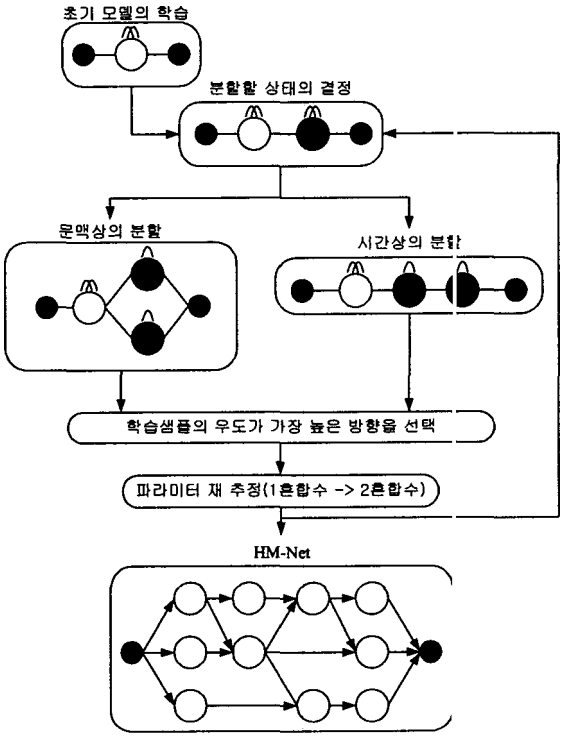

그림 1. SSS 알고리즘

일반적인 SSS 알고리즘을 그림 1에 나타내었다. 그림 1에 나타낸 일반적인 SSS 알고리즘에 음소결정트리를 접목한 예를 그림 2에 나타내었다.

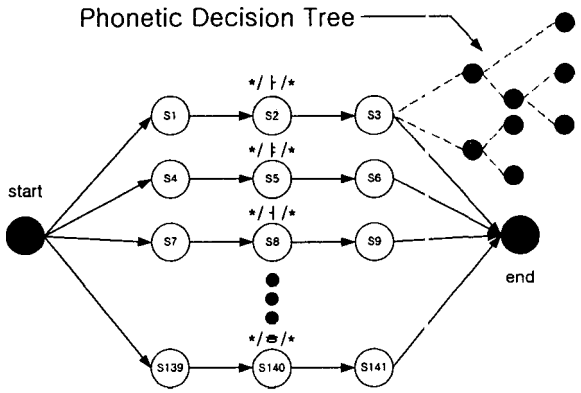

그림 2. HM-Net 초기 모델에 PDT를 접목한 예

PDT-SSS는 각 상태에 단일 정규 분포를 할당하여 두고, 상태 분할 시에는 새로운 분포를 구한다. 문맥 클래스는, 음소결정트리에 의해 2분할한다. SSS와의 차이점은 2혼합 분포를 2단일 분포로 분할하지 않고 질문 리스트에 의해 결정트리를 사용하여 문맥 상태를 분할하는 점이다. 이를 이용하면 SSS 알고리즘의 문제점을 해결 할 수 있다[11]. 즉, 음소결정트리의 장점인 미지의 문맥 요소를 간단히 모델링할 수 있으며 SSS 알고리즘의 장점인 정밀한 음향 모델을 작성할 수 있는 것이다.

3. 유사음소 단위의 고찰

대어휘 음성인식 시스템을 구현하기 위해서 면밀히 검토되어야 할 분야 중의 하나는 인식 단위에 관한 것으로 기본 인식 단위를 어떻게 설정하느냐에 따라 인식률의 차이를 보이게 된다[5]. 여기서는 문맥의존 음향 모델 작성법으로 도입된 HM-Net 모델 구성을 위한 새로운 유사음소 단위에 대해 고찰하기로 한다.

3.1 48 유사음소 단위

일반적으로 음성인식 시스템을 구현할 때 각 연구 그룹별로 기본적인 인식 단위를 정의하여 사용하고 있으며, 본 연구실에서는 표 1에 나타낸 것과 같이 48개의 유사음소 단위를 선정하여 사용하고 있다. 48 유사음소는 문맥독립 음향 모델을 작성할 때 기본 음소만으로 부족한 음성학적인 변이음을 추가하여 정의한 것이다[4]. 하지만 문맥의존 음향 모델인 HM-Net음향 모델은 훈련 데이터에 나타나는 수많은 선행 및 후행음소가 결합되어 다양한 종류의 변이음 모델이 자동 생성되기 때문에 기본 음소 단위에 변이음을 추가할 필요성이 없게 되어 문맥의존 음향 모델을 작성하기 위한 선행 및 후행음소의 중심 음소가 되는 기본 유사음소에서는 문맥독립인 경우 유효한 변이음을 고려할 필요가 없게 된다. 불필요한 기본 유사음소의 증가는 문맥의존 음향 모델 작성에서 부족한 학습 데이터의 훈련 효과를 분산시켜서 모델의 강건성을 저하시키는 원인이 된다.

표 1. 48 유사음소 단위

구분	48 유사음소 단위				
모음	aa /아/	axr /어/	ao /오/	uh /우/	U /으/
	ih /이/	ae /애/	eh /에/	ja /야/	jv /여/
	jo /요/	ju /유/	wa /와/	wv /워/	wE /외/
	we /웨,왜/	wi /위/	je /예,얘/	Wi /의/	
자음	b~ /ㅂ/	d~ /ㄷ/	g~ /ㄱ/	z~ /ㅈ/	hh~ /ㅎ/
	bb /ㅃ/	dd /ㄸ/	gg /ㄲ/	zz /ㅉ/	ss /ㅆ/
	s /ㅅ/	p /ㅍ/	t /ㅌ/	k /ㅋ/	ch /ㅊ/
	r /ㄹ/	n /ㄴ/	m /ㅁ/		
첫음절	b /ㅂ/	d /ㄷ/	g /ㄱ/	z /ㅈ/	hh /ㅎ/
종성	bl /ㅂ/	dl /ㄷ/	gl /ㄱ/	l /ㄹ/	ng /ㅇ/
묵음	sil				

3.2 유사음소 단위의 재정의

이와 같은 관점에서 현재 사용하고 있는 48 유사음소 중 첫음절에 사용되던 /b/, /d/, /g/, /hh/, /z/ 유사음소와, 종성으로 사용되던 /bl/, /dl/, /gl/, /l/ 유사음소 총 9개의 유사음소는 제외시키고 /b~/, /d~/, /g~/, /z~/, /hh~/ 유사음소들을 각각 /b/, /d/, /g/, /z/,

/hh/ 로 재정의 할 수 있다.

표 2. 48 유사음소와 39 유사음소의 문맥의존 음소 모델 /b/의 예

48 유사음소		39 유사음소
aa-b~+ao	aa-b+ao	aa-b+ao
aa-b1+b~	aa-b+b	aa-b+b
aa-b~+eh	aa-b+eh	aa-b+eh
aa-b1+eh	aa-b+eh	
aa-b1+g~	aa-b+g~	aa-b+g
aa-b1+hh~	aa-b+hh~	aa-b+hh
	...	
ae-b~+axr	ae-b+axr	ae-b+axr
ae-b~+jv	ae-b+jv	ae-b+jv
ao-b~+aa	ao-b+aa	ao-b+aa
ao-b1+aa	ao-b+aa	
ao-b1+g~	ao-b+g~	ao-b+g
ao-b~+ih	ao-b+ih	ao-b+ih
	...	
sil-b+aa	sil-b+aa	sil-b+aa
sil-b+ae	sil-b+ae	sil-b+ae
sil-b+ao	sil-b+ao	sil-b+ao
sil-b+axr	sil-b+axr	sil-b+axr
sil-b+eh	sil-b+eh	sil-b+eh
sil-b+ih	sil-b+ih	sil-b+ih
sil-b+jv	sil-b+jv	sil-b+jv
sil-b+uh	sil-b+uh	sil-b+uh
	...	
81가지		79가지

표 2는 국어공학센터(KLE)의 452단어 발성 리스트에서 발생할 수 있는 48 유사음소와 재 정의된 39 유사음소의 문맥의존 음향 모델 /b/에 대한 예를 나타내었다. 48 유사음소 단위에서 /b~/ 와 /b1/ 음소를 모두 /b/ 음소로 대체시키면 표 2의 중간 컬럼의 음소 구조로 나타난다. 이 경우 음영 부분의 문맥의존 유사음소에 주목하면 48 유사음소서 /aa-b~+eh/, /aa-b1+eh/ 유사음소가 39 음소에서는 /aa-b+eh/ 로 대표되어 하나의 문맥으로 사용되는 것을 볼 수 있다. 그리고 /ao-b~+aa/, /ao-b1+aa/ 음소의 경우도 /ao-b+aa/ 로 대표되고 있다. 여기서, /ao-b1+aa/는 /모음-자음(종성)+모음/의 VCV 형이다. 이 유형은 연음 현상에 해당되는 것으로 /ao-b1+aa/ 문맥이 /ao-b~+aa/ 으로 되는 것이므로 2개의 문맥을 /ao-b+aa/ 하나의 문맥으로 볼 수 있다. 이러한 연음 현상은 조음 현상과 더불어 오인식을 유발하는 대표적인 문제점으로 특히 연속 숫자음에서 많이 나타나는 현상이다. 또한, 48 유사음소 존재하는 2개의 문맥을 하나의 문맥으로 통합시키므로 해서 한정된 학습 데이터를 이용할 경우 2배의 학습 효과를 얻을 수 있는 부수적인 이점도 있다.

표 3. 48 유사음소와 39 유사음소의 비교

48 유사음소	비고	39 유사음소	비고
g, d, b, z, hh	첫음절 초성	g, d, b, z, hh	초, 종성
g~, d~, b~, z~, hh~	초성		
g1, d1, b1	종성		
r	초성	r	초, 종성
l	종성		

48유사음소 중 첫 음절에 나타나는 유사음소들은 표 2와 같이 음소의 재 정의된 후에도 표기상으로는 아무런 변화가 없게 된다. 이와 같은 방법으로 48 유사음소의 /d/, /g/, /z/, /h/, /r/ 계열은 /b/ 계열의 음소와 같은 경우로 취급하여 총 39 유사음소로 재정의된다.

4. 인식 실험과 결과 및 고찰

본 논문에서 도입한 HM-Net 모델의 구조 결정법과 재 정의한 39 유사음소 단위의 유효성을 확인하기 위해 단어, 숫자음, 연속음성에 대해 각각 인식 실험을 수행하였다. 여기서 사용된 음성 데이터베이스는 표 4에 나타낸 것과 같이 고립 단어 인식에는 국어공학센터(KLE) 452 단어를, 숫자음 인식에는 KLE 4연속 숫자음을, 연속 음성 실험에는 한국과학기술원(KAIST)에서 수집한 무역 상담용 연속 음성 데이터베이스를 각각 사용하였다. 음성인식 알고리즘은 Word-pair 문법과 N-gram 언어 모델을 인식 문법으로 하는 One-Pass Viterbi 알고리즘을 사용하였다. 사용한 음성 데이터의 분석 조건은 표 5와 같다.

표 4. 음성 데이터베이스

실험	단어 인식	숫자음 인식	연속 음성인식
사용 데이터	KLE 452 단어	KLE 4연속 숫자음	무역 상담용 연속 음성
학 습	35명 1회 발성	35명 4회 발성	90명
평 가	3명 2회 발성	3명 4회 발성	10명
인식형태	화자 독립	화자 독립	화자 독립

표 5. 음성 데이터의 분석 조건

주파수	16kHz
양자화	16bit
프레임 길이	25ms
프레임 주기	10ms
분석창	Hamming Windows
특징 파라미터	12차 LPC-MEL cepstrum + delta power + 1, 2차의 회귀 계수 = 39차원

4.1 고립 단어 인식 결과

고립 단어 인식 실험의 경우 HM-Net의 상태수를 정밀하게 하여 상태 분할을 수행하였다. 상태 분할 수의 증가에 따른 인식률의 변화 정도를 파악하기 위해서 상태수를 200에서 3000까지 200 상태씩 증가시키면서 48 유사음소와 39 유사음소 인식률을 구하였다. 표 6에 고립 단어 인식률을 나타내었다. 표 6에서 알 수 있듯이 48 유사음소 경우 상태수 600에서 98.71%로서 최고 단어 인식률을 얻었으나 그 이후의 상태에서는 다소 감소하는 결과를 보이고 있다. 반면에 39 유사음소 단위의 단어 인식률은 상태수 2200일 때까지 인식률의 증가를 보이며, 그 후 조금 감소하는 것을 확인할 수 있다. 48 유사음소 단위의 경우 문맥독립 모델에서 사용하면 음소를 그대로 사용하였기 때문에 상태가 낮을 때는 상태 분할을 충분히 수행하지 못한 단계이므로 문맥독립 모델에 가깝다고 볼 수 있다. 하지만 상태 분할이 충분히 진행되면 문맥의존형 모델에 가까워지고 상태간의 공유도 좀 더 정밀하게 수행된다. 따라서 39 유사음소 단위가 문맥의존형의 기본 인식 단위에 유효함을 알 수 있다.

표 6. 48, 39 유사음소의 고립 단어 인식률(%)

상태수	200	400	600	800	1000	1200
48 유사음소	97.27	98.64	98.71	98.56	98.41	98.41
39 유사음소	96.57	98.27	98.60	99.08	99.08	99.12
상태수	1400	1600	1800	2000	2200	2400
48 유사음소	98.30	98.30	98.16	98.19	98.38	98.30
39 유사음소	99.04	99.04	99.19	99.15	99.26	99.12
상태수	2600	2800	3000			
48 유사음소	98.19	97.94	97.86			
39 유사음소	99.08	99.08	99.08			

4.2 4연속 숫자음 인식 결과

48 유사음소 단위와 재 정의한 39 유사음소 단위의 유효성을 확인하기 위해서 HM-Net 모델의 상태수를 200에서 700까지 100 상태씩 증가시키면서 혼합수가 4개인 HM-Net 문맥의존 음향 모델을 작성한 후 4연속 숫자음 인식 실험을 수행하였다. KLE의 숫자음에 대한 음성 데이터의 양이 적어 큰 상태수를 가진 모델의 파라미터를 추정하는데 어려움이 있어 상태수를 700으로 설정하였다. 숫자음 인식 실험 결과를 표 7에 나타내었다.

표 7. 48, 39 유사음소의 4연속 숫자음 인식률(%)

상태수	200	300	400	500	600	700
48 유사음소	81.43	83.33	85.95	89.76	91.90	91.43
39 유사음소	90.48	94.05	92.62	95.24	95.00	95.71
인식률 차이	9.05	10.72	6.67	5.48	3.10	4.28

표 7에서 재 정의한 39 유사음소가 기존의 48 유사음소 단위에 비해 전체적으로 인식 성능이 더 향상됨을 알 수 있으며, 최고 인식률을 나타낸 상태수 700일 때 39 유사음소 단위의 경우 95.71%로서 48 유사음소 단위에 비해 평균 4.28% 향상됨을 알 수 있었다. 이상의 숫자음 인식 실험을 통하여, 재 정의한 39 유사음소 단위가 연음 현상이 많은 숫자음에서 유효함을 확인할 수 있다.

4.3 연속 음성인식 결과

본 논문에서 도입한 HM-Net 모델의 구조 결정법과 재 정의한 유사음소 단위의 유효성을 확인하기 위해 남성 10명이 발성한 KAIST의 무역 상담 연속 음성을 대상으로 연속 음성인식 실험을 수행하였다. HM-Net 문맥의존 음향 모델은 혼합수 4에 상태수가 각각 1000, 2000, 3000을 가지도록 작성하였다. 언어 모델은 N-gram 모델로서 N이 2인 2-gram 모델을 사용하였으며, 인식 방법은 One-Pass Viterbi 알고리즘을 사용하였다. 인식 실험 결과를 표 8에 나타내었다.

표 8. 48, 39 유사음소의 연속 음성 문장 인식률(%)

상태수	1000	2000	3000
48 유사음소	73.16	73.40	70.05
39 유사음소	84.81	88.69	88.26
인식률 차이	11.75	15.29	18.21

표 8에 나타낸 것과 같이 연속 음성인식 실험에서도 39 유사음소 단위가 48 유사음소 단위에 비해 각 상태수에 따라 11.75%, 15.29%, 18.21%의 인식률 향상이 있음을 확인할 수 있다. 이와 같은 인식률 향상은 연음 현상이나 조음 현상이 가장 빈번히 나타나는 연속 음성에서 39 유사음소 단위가 48 유사음소 단위에 비해 우수한 성능을 보임으로써 본 논문에서 도입한 HM-Net 구조 결정법을 적용한 문맥의존 음향 모델에 재 정의한 39 유사음소 단위가 유효함을 확인할 수 있었다.

5. 결 론

본 논문에서는 음소결정트리 기반 HM-Net 모델의 구조 결정법을 도입하여 작성한 문맥의존 음향 모델에 적합한 유사음소 단위를 재정의 하였다. 이에 대한 유효성을 확인하기 위해 인식 실험을 수행한 결과 제안한 39 유사음소 모델을 사용한 경우 48 유사음소모델에 비해 고립 단어 인식의 경우 평균 0.61%, 4연속 숫자음인식의 경우 평균 6.55%, 연속 음성인식의 경우 평균 15.08%의 인식률 향상이 있었다. 따라서 본 논문에서 재 정의한 유사음소 단위가 문맥의존 HM-Net 음향 모델에 유효함을 확인할 수 있었다.

※ 본 연구는 한국과학재단 목적기초연구 (R01-2000-000-00276-0) 지원으로 수행되었음.

참 고 문 헌

[1] 임영춘, 오세진, 김범국, 정현열, "HMnet을 이용한 한국어 음소인식에 관한 연구," 한국음향학회 영남지회, 2000.

[2] 김선일, 홍기원, 이행세, "국어 종성 자음의 음향학적 특징에 관한 연구," 한국음향학회지, 제14권 1호, pp. 65-72

[3] 최인정, 권오욱, 박종렬, 박용규, 김도영, 정호영, 은종관, "대용량 한국어 연속음성인식 시스템 개발," 한국음향학회지, 제14권 제5호, pp. 44-50, 1995.

[4] 김유진, 김회린, 정재호, "인식 단위로서의 한국어 음절에 관한 연구," 한국음향학회지, 제16권 제3호, pp. 64-72, 1997.

[5] 김호경, 구명환, "기본음소 설정을 위한 음소인식률 이용 방안 연구," 제15회 음성통신 및 신호처리 워크샵 논문집, pp. 328-331, 1998.

[6] 이승훈, 김회린, "가변어휘 음성인식기의 음향모델 개선 및 성능분석," 한국음향학회지, 제18권 제8호, pp. 3-8, 1999.

[7] 박현상, 은종관, 박용규, 권오욱, "Diphone 단위의 hidden Markov model을 이용한 한국어 단어인식," 한국음향학회지, 제13권 제1호, pp. 14-23, 1994.

[8] 박준영, "한국어 단어인식을 위한 최적의 연속 HMM모델에 관한 연구," 석사 학위 논문, 영남대학교, 1995.

[9] J. Takami, and S. Sagayama, "A successive state splitting algorithm for efficient allophone modeling," Proc. of ICASSP'92, Vol. 1, pp. 573-576, 1992.

[10] Ostendoft, and H. Singer, "HMM topology design using maximum likelihood successive state splitting," Computer Speech and Language, Vol. 11, pp. 17-41, 1997.

[11] Se-Jin Oh, Chul-Jun Hwang, Bum-Koog Kim, Hyun-Yeol Chung, and Akinori Ito, "New state clustering of hidden Markov network with Korean phonological rules for speech recognition," IEEE 4th workshop on Multimedia Signal Processing, pp. 39-44, 2001.