강건한 한국어 연속음성인식을 위한 유사음소단위에 대한 연구

서 준 배, 김 주 곤, 김 민 정, 정 호 열, 정 현 열
영남대학교 전자정보공학부

A Study on PLU (Phone-Likely Unit) for Korean Continuous Speech Recognition

Jun-Bae Seo, Joo-Gon Kim, Min-Jung Kim, Ho-Youl Jung, Hyun-Yeol Chung
Dept.of Information and Communication Eng., Yeungnam University
{sjbda,speech,manjuk}@yumail.ac.kr {hoyoul,hychung}@yu,ac,kr

요약

본 논문은 한국어 연속음성인식에 효율적인 문맥의존 음향모델 수에 대한 연구로써 유사음소단위 수에 따른 인식 성능을 비교, 평가 하였다. 기존에 본연구실에서는 48음소를 기본인식단위로 이용하고 있으나 연속음성인식의 경우 문맥종속모델이 사용되고 문맥종속모델은 변이음을 고려한 음소가 이미 포함되어 있어 이를 고려하면 기본 음소를 줄이므로서 계산량의 감소와 인식 성능 향상을 기대할 수 있을 것으로 생각된다. 따라서, 본 논문에서는 기존의 48음소와 이를 39음소로 줄여 인식실험에 사용하여 그 성능을 비교 평가하기로 하였다. 이를 위하여 다양한 태스크의 데이터베이스를 통합하여 부족한 문맥요소들을 확장한 후 인식실험을 수행하였다. 실험결과 변이음의 개수를 줄이면서도 인식 성능저하가 없음을 확인할 수 있었으며 연속 음성의 경우 39음소를 이용한 경우가 10%정도의 향상된 인식성능을 얻을 수 있음을 확인할 수 있었다.

I. 서 론

음성인식 시스템은 1970년대 초부터 활발히 연구되어 왔으며, 대표적인 인식기법으로는 음성발성 시간상에서의 패턴정합에 의해 음성을 인식하는 동적프로그래밍(Dynamic Programmig;DP)정합 방법, 인식 계산량과 메모리를 적게 하기 위한 데이터 압축 기술을 이용한 벡터 양자화(Vector Quantization;VQ) 기법, Markov 모델의 확률적 추정에 의한 기법을 도입한 은닉마르코프모델(Hidden Markov Model;HMM)과 음성의 인지 과정을 모델화한 인공신경 회로망 등을 이용한 것 등이 있다[1]. 음성 인식 성능은 앞에서 언급한 패턴 인식 방법들 외에, 표준 패턴으로 저장하는 음성인식 단위에 따라 그 성능이 크게 좌우된다[2]. 앞의 인식 방법들이 언어에 관계없이 적용될 수 있는 기법들임을 고려할 때, 언어에 따른 인식성능의 차이는 어떠한 인식단위를 이용하느냐에 종속된다고 할 수 있다[3][4][5].

본 논문의 구성은 다음과 같다. 2장에서 연속음성인식을 위한 기본 인식단위를 정의하고 3장에서 실험에 사용된 대용량 음성데이터베이스에 대해서 살펴본다. 4장에서는 인식실험을 통해 유사음소 단위별 인식결과를 검토하고, 마지막으로 5장에서 결론을 맺도록 한다.

II. 연속음성인식을 위한 유사음소단위

최근 음성인식에서 널리 사용되고 있는 인식 단위로는 음소, 음절, 단어 등의 언어적으로 정의된 단위들과 유사음소단위(Phoneme Likely Unit; PLU)나 음향학적인 우도에 근거한 단위들이 사용되고 있다[1][6]. 특히, 대 어휘 연속음성 인식기에서의 기본 인식 단위는 음소이다. 음소는 단어나 음절에 비해 그 수가 작고 학습에 필요한 충분한 자료를 모으기가 용이하다는 장점이 있다[2]. 그러나 음소는 좌우에 위치하는 음소에 영향을 많이 받으므로, 이를 고려하여 세분화된 문맥의존 음소모델이 구성되어야 한다. 이전 실험들에서 문맥독립 음소는 문맥의존 음소에 비해 많은 변이를 포함하므로 모델링이 어려워지고 인식률에 있어서도 저조한 결과를 보였었다[6]. 따라서 문맥 독립 음소를 사용할 경우 단위모델에 대한 정확한 모델링뿐만 아니라 분별학습, 후처리 등의 충분한 뒷받침 없이는 높은 인식률을 기대하기 어렵다[9]. 반면, 문맥의

존 음소모델은 문맥독립 모델에 비해 음향의 가지 수는 많지만 음소에 의한 변이음을 고려한 모델[10][11]로서 강건한 음향모델을 생성하는 방법으로 많은 연구가 진행되고 있다. 본 논문에서는 음향모델생성 방법 중에 강건한 음향모델을 생성하기 위하여 은닉 마르코프 네트워크(Hidden Markov Network;HM-Net)을 적용하였다. HM-Net은 HMM의 상태를 정해진 상태 모델링 방식에서 연속상태분할(Successive State Split;SSS)알고리즘을 적용하여 음향학적 정보에 따라 자동으로 상태 분할하는 음향 모델링 방법이다. 이러한 문맥 의존 음향모델링 방법을 대어휘 연속음성인식에 적용하며, 대어휘 연속음성인식을 위해 고려해야할 점들을 검토한다.

연속음성인식에서는 최적의 인식단위 선정과 충분한 학습 데이터를 고려해야 한다. 일반적으로 연속음성인식에서는 다양한 변이음들의 음향적 변이를 충분히 표현할 수 있는 트라이폰을 사용하므로 충분한 학습을 위해서는 대용량의 학습 데이터가 요구된다. 또, 인식의 기본단위로서 기존의 문맥독립모델에서 사용된 음소간의 변이정보를 포함한 48유사음소단위와 변이정보를 제외시켜 음소단위에 가깝게 재 정의한 39유사음소단위를 기준으로 각각 문맥의존 음향 모델을 작성하여 최적의 인식단위를 고려할 필요가 있다[12]. 이러한 두 가지 문제점을 보완하고 해결하기 위해서는 대용량 데이터베이스로 구축하는 것이 가장 이상적이나 이러한 음성 데이터베이스를 구축하는 것은 현실적으로 불가능하다. 따라서, 본 논문에서는 기존의 음성 데이터베이스들을 통합하여 다양한 환경과 다양한 문맥구조를 가진 대용량 데이터베이스로 구성하였다. 이러한 대용량 데이터베이스를 사용하여 문맥의존 음향모델 작성방법인 HM-Net으로 대어휘에 적합한 음소 수에 대해 검토하고자 한다.

유사음소 단위는 최소 인식단위로 많이 사용되며 기본적인 음소에 변이음을 포함하고 있는 음소이다. 음향학적 및 음성학적 유사성이 큰 경우에는 음소와 유사음소 단위는 동일하게 취급될 수 있지만 그렇지 않을 경우 큰 차이가 있다. 48유사음소 단위는 자음의 경우 동일음소이지만 음향학적 음가가 다르기 때문에 이를 구분시켜 정의된 유사음소 단위이며, 39유사음소 단위는 변이음을 포함하지 않는 음소정의에 가까운 유사음소 단위이다. 대어휘 연속음성인식에서는 보다 높은 인식성능을 위해서 음소의 좌우 문맥 정보를 사용하기 때문에 이미 음성의 변이정보를 모두 포함하게 된다. 이러한 점을 고려하여 변이음 정보였던 음소를 제외시킨 것이 39 유사음소 단위이다. 표 1에 기존의 48유사음소와 재 정의된 39음소에 대해 나타낸다. 39유사음소 단위는 음성데이터의 부족한 학습데이터의 훈련효과를 분산시키는 것을 줄임으로 한정된 데이터에서 효율적인 학습효과를 가져왔다[6].

구분	48 유사음소단위				
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				

표 1. 48(짙은 부분 제외시 39 유사음소)유사음소단위

III. 대어휘 연속음성인식을 위한 대용량 데이터베이스 구성

음성인식의 실용화를 위해서는 다양한 문맥구조 및 태스크의 환경을 포함하는 대용량 데이터베이스가 필요하지만 현실적으로는 어려운 실정이다. 따라서, 본 논문에서는 대용량 데이터베이스를 구성하기 위하여 기존의 다양한 환경 및 문맥구조를 가지는 여러 데이터베이스를 통합하여 대용량 데이터베이스를 구성하였으며, 이를 표 2에 정리하였다.

사용된 데이터베이스는 국어 공학센터에서 구축한 숫자, 연속숫자, PBS(Phonetic Balance Sentence), PBW(Phonetic Balance Word) 데이터베이스, ETRI에서 제공하고 있는 445단어로 구성된 PBW와 1000명의 화자로 각 100개의 단어와 숫자로 구성된 대규모 데이터베이스를 이용하여 다양한 태스크를 포함하는 대용량 음성 데이터베이스를 구축하였다. 또한, 연속문장 데이터베이스를 위해서 한국과학기술원에서 작성한 무역상담용 연속음성 데이터베이스를 대용량 데이터베이스에 포함 시켰다. 이렇게 함으로서 특정 태스크에 적합하도록 구성되어 저있는 기존 데이터베이스를 이용하여 숫자음, 연속 숫자음, 일상생활에 사용되는 단어와 음소별로 잘 구성된 단어(PBW), 문장 데이터 등이 잘 분포된 대용량 데이터베이스를 구축할 수 있었다.

학습에 사용된 화자수는 2,370명이며 학습된 단어 및 문장 어휘수는 총 281,920개로서 기존의 음성데이터를 최대한 통합하였다.

DB종류	어휘수 (문장)	전체 DB			비고
		화자수	반복횟수 /어휘	총단어 수(문장)	
KLE PBW452	452	남자35/3명 여자32명	70회	31640	훈련 /인식
KLE PBS	600	남자9/1명 여자9/1명	1회	1200	훈련 /인식
KLE Digit	41	남자37명 여자33명	4회	11480	훈련
KLE C-Digit	35	남자37명 여자33명	4회	9800	훈련
ETRI PBW445	445	남자21명 여자19명	40회	17800	인식
TRADE DB	100	90/10명	1회	10000	훈련 /인식
ETRI 단어	10000	1000	10회	100000	훈련
ETRI 숫자	10000	1000	10회	100000	훈련
합계	21673	2370		281920	

표 2.음성 데이터 목록

IV. 인식 실험 및 고찰

다양한 태스크(단어, 숫자, 연속문장)의 음성데이터로 학습한 모델을 이용하여 어휘독립 연속음성 인식실험을 수행하였다. 학습에 사용된 음성데이터는 국어공학연구소에서 작성한 PBW 452단어, 숫자음, 4연속 숫자음, PBS , ETRI에서 PC환경에 중급마이크로 녹음된 단어 1,000명분, 숫자 1,000명분 데이터와 KAIST에서 작성한 무역상담용 DB를 사용하였고 인식에 사용된 음성 데이터는 KAIST무역 상담용 DB의 학습에 사용되지 않는 문장으로 하였다.

문맥의존음향모델에서 부족한 음성데이터로 인해 발생한 미지의 문맥요소 및 다양한 변이음들을 대용량 음성 코퍼스를 이용하여 보완하였다.

4.1 어휘독립 대어휘 연속음성인식 실험

어휘독립 연속음성인식 실험에서 단어 인식률과 문장인식률을 그림 1,2에 각각 나타낸다. 단어인식률의 경우 최고 17.5%의 성능차를 보였으며, 문장인식률의 경우 최고 30%의 성능차를 보였다. 이 실험 결과로부터 39음소가 연음 및 변이음이 자주 발생하는 어휘독립 연속음성 인식환경에서 48음소보다 효율적인 음소 체계임을 인식 성능을 통해 확인할 수 있었다.

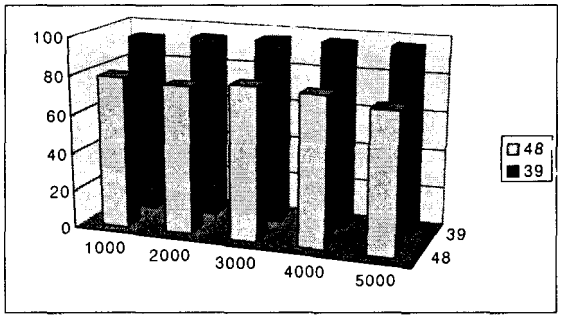

그림 1. 어휘독립 단어인식률

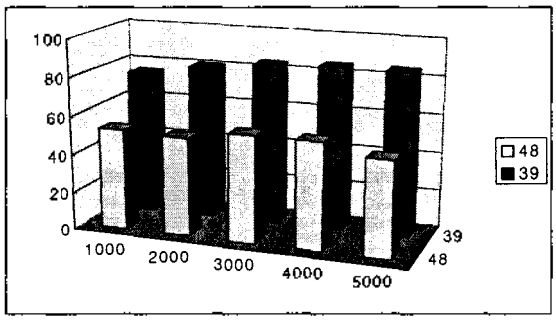

그림 2. 어휘독립 문장인식률

	state / PLU	1000	2000	3000	4000	5000
Word	39	93.56	95.00	96.12	96.78	96.55
	48	78.91	76.73	79.28	77.75	73.14
Sentence	39	75.45	80.63	84.17	85.74	85.43
	48	52.43	51.36	55.70	55.48	49.90

표 3. 어휘독립 연속음성인식률

4.2 어휘종속 대어휘 연속음성인식 실험

어휘종속 대어휘 연속문장인식 실험에서 단어 인식률과 문장인식률을 그림 3과 4에 각각 나타내며, 표 4에 함께 나타낸다. 학습에 포함된 어휘를 포함한 PBS문장을 인식한 결과, 단어인식률의 경우 최대 1.65%이며 문장인식률에서는 최대 2.89%의 인식률 차이를 확인할 수 있었다. 이 실험결과에서 어휘 종속의 경우 어휘독립 연속 음성인식의 경우와 같이 급격한 인식률 차이는 없지만 39음소가 48 음소 보다 향상된 인식 성능을 보였다.

이러한 실험결과에서 어휘 종속 및 어휘독립 연속 음성인식, 모두에서 39음소가 48음소 보다 향상된 인식 성능을 보여 39음소가 효율적인 음소체계임을 확인할 수 있었다.

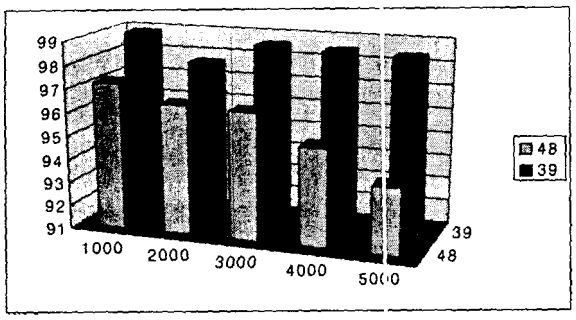

그림 3. 어휘종속 단어 인식률

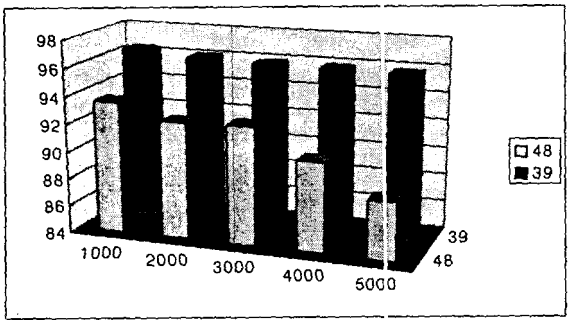

그림 4. 어휘종속 문장인식률

	state / PLU	1000	2000	3000	4000	5000
Word	39	98.87	97.70	98.38	98.48	98.40
	48	97.22	96.48	96.39	95.16	93.81
Sentence	39	96.42	96.10	95.91	95.85	95.81
	48	93.53	92.42	92.35	90.46	88.15

표 4. 어휘종속 연속음성인식률

V. 결론

본 논문은 한국어 연속음성인식에 효율적인 문맥의존 음향모델 수에 대한 연구로써 유사음소단위 수에 따른 인식 성능을 비교, 평가 하였다. 연속음성인식에 이용되는 문맥종속모델의 경우 변이음을 고려하여 모델이 작성되므로 이를 고려하여 기본 음소 수를 48음소로부터 39음소로 줄일 수 있었다. 39음소를 이용한 경우와 인식 성능 평가를 위하여 48음소를 이용한 경우와 비교 평가를 수행하였다. 비교평가를 위한 인식엔진은 HM-Net을 이용하였다. 또한, 강건한 문맥의존 음향모델을 생성하기 위해서 다양한 태스크의 음성데이터베이스를 통합하여 대용량 데이터베이스를 구성하였으며, 대용량 데이터베이스로 각 음소별 HM-Net음향모델을 학습시켜 대어휘 연속음성인식 실험을 수행하였다.

실험결과, 어휘독립 연속음성인식에서 단어인식률은 17%이상, 문장인식률은 30%의 인식성능향상을 보였으며, 학습에 포함된 KLE PBS문장 데이터의 경우에도 39유사음소단위가 단어 인식률은 1.65%, 문장인식률은 2.89%의 성능향상을 보여, 39음소가 발음변이가 빈번히 일어나는 연속음성인식 환경에서는 48음소보다 효과적임을 알 수 있었다.

참 고 문 헌

[1] 황영수, "한국어 인식을 위한 인식단위와 학습 데이터 분류 방법에 대한 연구," 한국 신호처리시스템학회 논문지, 제4권 제2호, pp. 40-45, 2003.

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

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

[4] Kai-Fu Lee, Hsiao-Wuen Hon, "Large-vocaulary speaker-Independent Continuous speech recognition Using HMM," ICASSP. pp. 749-752, 1990.

[5] S.Kanthak, H. Ney, " Multilingual Acoustic Modeling Using Graphemes," ECSCT, Vol 2, pp. 1145-1148, 2003.

[6] 임영춘, " HM Net 유사음소 단위에 관한 연구," , 2002.

[7] Julian James Odell, "The Use of Context in Large Vocabulary Speech Recognition," University of Cambridge Ph.D, 1995.

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

[9] M. Suzuki, S. Makino, A. Ito, H. Aso, and H. Shimodaira, "A new HMnet construction algorithm requiring no contextual factors," IEICE Trans. Info. & Syst., Vol. E78-D, No. 6, pp. 662-669, 1995.

[10] Rubem Dutra Ribeiro Fagundes, Juarez Sagebin Correa, Pierre Dumouchel, " A New phonetic model for continuous speech recognition systems", ICSP'02 Procedings, pp. 572-575, 2002

[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.

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