잡음환경하의 연속 음성인식을 위한 유사음소단위 분석

신 광호, 임 수호, 서 준배, 김 주곤, 정 호열, 정 현열
영남대학교 정보통신공학과

An Analysis on Phone-Like Units for Korean Continuous Speech Recognition in Noisy Environments

Guang-Hu Shen, Soo-Ho Lim, Jun-Bae Seo, Joo-Gon Kim, Ho-Youl Jung, Hyun-Yeol Chung
Dept.of Information and Communication Eng., Yeungnam University
E-mail: guanghosin@yumail.ac.kr

요 약

본 논문은 잡음환경 하에서의 효율적인 문맥의존 음향모델 구성에 대한 기초연구로서 잡음환경 하에서의 유사음소단위 수에 따른 연속 음성인식 성능을 비교, 평가한 결과에 대한 보고이다. 기존의 연구[1,2]로부터 연속음성인식의 경우 문맥종속모델은 변이음을 고려한 39유사음소를 이용한 경우가 48유사음소를 이용하는 것보다 더 좋은 인식성능을 나타냄을 알 수 있었다. 이 연구 결과를 바탕으로 본 연구에서는 잡음환경에서도 효율적인 문맥의존 음향모델을 구성하기 위한 기초 연구를 수행하였다. 다양한 잡음환경을 고려하기 위해 White, Pink, LAB 잡음을 신호 대 잡음비(Signal to Noise Ratio) 5dB, 10dB, 15dB 레벨로 음성에 부가한 후 각 유사음소단위 수에 따른 연속음성인식 실험을 수행하였다. 그 결과, 39유사음소를 이용한 경우가 48유사음소를 이용한 경우보다 clear 환경인 경우에 약 7%와 17% 향상된 단어인식률과 문장인식률을 얻을 수 있었으며, 각 잡음환경에서도 39유사음소를 이용한 경우가 48유사음소를 이용한 경우보다 평균적으로 17%와 28% 향상된 단어인식률과 문장인식률을 얻을 수 있어 39유사음소 단위가 한국어 연속음성인식에 더 적합하고 잡음환경에서도 유효함을 확인할 수 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]. 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. 특히, 대 어휘 연속음성 인식기에서의 인식단위는 음소이다. 음소는 단어나 음절에 비해 그 수가 적고 학습에 필요한 충분한 자료를 모으기가 용이하다는 장점이 있다[1,2]. 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[2]. 따라서 문맥 독립 음소를 사용할 경우 단위모델에 대한 정확한 모델링뿐만 아니라 분별학습, 후처리 등의 충분한 뒷받침 없이는 높은 인식률을 기대하기 어렵다[6]. 반면, 문맥의존 음소모델은 문맥독립 모델에 비해 음향의 가지 수는 많지만 음소에 의한 변이음을 고려한 모델[7]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)를 적용하였다. HM-Net은 HMM의 상태를 정해진 상태 모델링 방식에서 연쇄상태분할(Successive State Split; SSS) 알고리즘을 적용하여 음향학적 정보에 따라 자동으로 상태를 분할하는 음향 모델링 방법이다. 이러한 문맥의존 음향모델링 방법을 연속음성인식에 적용하여, 연속음성인식에서 고려해야할 점들을 검토한다.

실제 언어 환경은 매우 다양한 형태로 나타나는 잡음환경의 영향을 받게 된다. 그러므로 연속음성인식에서는 최적의 인식단위 선정뿐만 아니라 잡음환경을 고려한 음

성 데이터베이스를 사용할 필요가 있다[9]. 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[5]. 따라서 학습 데이터에 잡음 환경요인을 고려하여, 3가지 잡음(White, Pink, LAB)을 신호 대 잡음비(Signal to Noise Ratio) 5dB, 10dB, 15dB 레벨로 음성에 부가하여 연속음성 데이터베이스로 사용하도록 했다. 이러한 다양한 환경을 고려한 데이터베이스를 사용하여 문맥의존 음향모델 작성방법인 HM-Net으로 연속음성인식에 적합한 음소 수에 대해 검토하고자 한다.

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

2. 연속음성인식을 위한 기본 인식단위

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 경우에는 큰 차이가 있다. 48 유사음소는 문맥독립 음향 모델을 작성할 때 기본 음소만으로는 부족한 음성학적인 변이음을 추가하여 정의한 것이다[8]. 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 된다. 그러므로 문맥의존 음향 모델을 작성하기 위한 선행 및 후행음소의 중심 음소가 되는 기본 유사음소에서는 문맥독립인 경우에서 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.

39유사음소 단위는 변이음을 포함하지 않는 음소정의에 가까운 유사음소 단위이다. 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 1은 기존의 48유사음소에 대해 나타낸다. 48유사음소의 /d/, /g/, /z/, /h/, /r/ 계열은 표 2에서와 같은 경우로 취급하여 총 39유사음소로 재 정의된다. 39유사음소 단위는 음성데이터의 부족한 학습데이터의 훈련효과를 분산시키는 것을 줄일 수 있다. 그러므로 제한된 학습 데이터에서 더 많은 학습 데이터를 확보하게 되어 좀 더 강건한 모델을 학습하게 된다[1,2].

표 1. 48유사음소단위

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 2. 48유사음소와 39유사음소의 비교

48 유사음소	비교	39 유사음소	비교
g, d, b, z, hh	첫음절 초성	g, d, b, z, hh	초, 종성
g~, d~, b~, z~, hh~	초성		
gl, dl, bl	종성		
r	초성	r	초, 종성
l	종성		

3. 잡음환경에서의 음성 표현

잡음은 보통 백색잡음(White noise)과 유색잡음(Colored noise)으로 구분된다. White 잡음은 스펙트럼이 모든 주파수대역에서 균일하며, 시간영역에서 샘플값이 서로 상관성이 없다. 이와 달리 Pink 잡음은 주파수 대역에서의 스펙트럼이 일정하지 않은 유색잡음이며, 다양한 분야에서 흔히 접할 수 있다. LAB잡음은 연구실환경에서 녹음된 배경잡음이며, 이 잡음은 기계적인 잡음, 사람들이 주위에서의 대화소리 등 여러 가지 배경잡음이 포함되고 있다. 배경잡음에서 기계적인 원인에 의해 발생되는 저주파

잡음들은 전력 스펙트럼을 감소시키는 경향이 있으며, 주변 환경 잡음에 의해 발생되는 스펙트럼은 높은 주파수 특성을 가지며, 일반적으로 불규칙한 특성을 가진다[9].

평균이 영인 신호를 가정할 때, 신호 대 잡음비는 식(1)과 같이 정의될 수 있다.

$$SNR = 10\log\frac{E_s}{E_n} \quad (1)$$

여기서 E_s와 E_n은 음성신호와 잡음의 평균 에너지이다. 입력 음성신호를 $s(t)$, 선형 시불변 필터를 $h(t)$, 부가잡음을 $n(t)$로 하면, 열화된 음성신호 $x(t)$는 시간영역에서 식(2)과 같이 표현할 수 있다.

$$x(t) = s(t) * h(t) + n(t) \quad (2)$$

본 논문에서는 다양한 잡음환경을 고려하기 위해 White, Pink, LAB 잡음을 신호 대 잡음비 5dB, 10dB, 15dB 레벨로 음성에 부가한 후 각 유사음소단위 수에 따른 연속음성인식 실험을 수행하였다.

4. 인식 실험 및 고찰

본 논문에서 사용한 음성 데이터는 KAIST무역 상담용 DB이다. 잡음환경을 고려하기 위해 White, Pink, LAB잡음을 깨끗한 음성 데이터에 부가하였으며, 발성화자 중100명분에서 90명분을 학습데이터로 이용하였고, 나머지 10명분으로 화자독립 인식실험을 수행하여 유효성을 비교검토하였다. 인식을 위한 음향모델은 2000상태 8혼합수의 HMNet모델을 이용하였으며, 음성인식 알고리즘은 Word-pair 문법을 인식 문법으로 하는 One-Pass Viterbi 알고리즘을 사용하였다. 사용한 음성 데이터의 분석조건은 표 3과 같다.

표 3. 음성 데이터의 분석조건

주파수	8kHz
양자화	16bit
프레임 길이	25ms
프레임 주기	10ms
분석창	Hamming Windows
특징 파라미터	12차 LPC-MEL cepstrum+delta power + 1,2차의 회귀 계수 = 39차원

1) 무잡음 환경하의 연속음성인식 실험

무잡음환경하의 어휘독립 연속음성인식 실험에서 단어인식률과 문장인식률을 표 4에서 나타낸다. 단어인식률의 경우 약 7%의 성능차를 보였으며, 문장인식률의 경우 약 17%의 성능차를 보였다. 이 실험 결과로부터 39음소가 연음현상 및 변이음현상이 자주 발생하는 연속음성인식 환경에서 오인식을 유발하는 문제를 48음소보다 더 효율적으로 처리할 수 있음을 알 수 있었다. 연속음성인식에서 39음소가 48음소에 비해 더 적합한 음소 체계임을 인식 성능을 통해 확인할 수 있었다.

표 4. 어휘 독립 연속음성인식률

인식률 \ PLU	39	48
단 어	97.69	90.64
문 장	88.83	71.56

2) 잡음환경하의 연속음성인식 실험

White, Pink, LAB 잡음을 신호 대 잡음비 5dB, 10dB, 15dB 레벨로 깨끗한 음성에 부가한 후 각 유사음소단위 수에 따른 연속음성인식 실험을 수행 하였다. 인식결과는 그림 1,2,3에 나타 내였다.

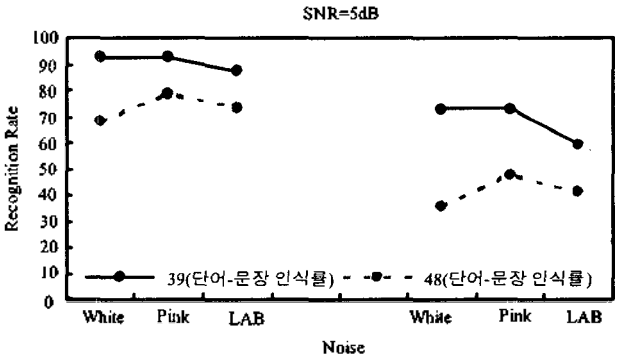

그림 1. 신호 대 잡음비 5dB일 때 39음소와 48음소의 잡음별 단어 및 문장인식률

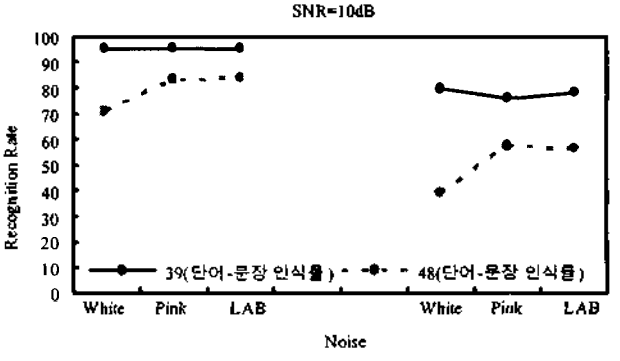

그림 2. 10dB(SNR)일 때 39음소와 48음소의 잡음별 단어 및 문장인식률

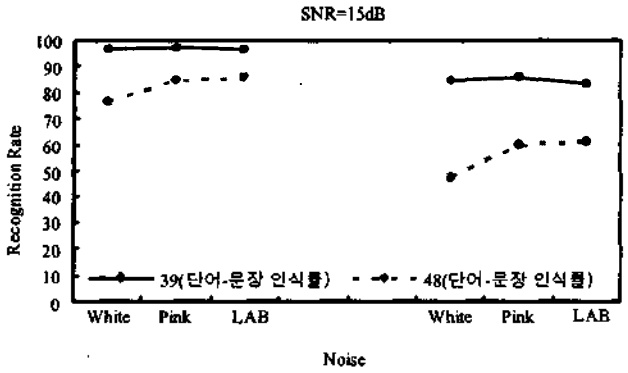

그림 3. 15dB(SNR)일 때 39음소와 48음소의 잡음별 단 및 문장인식률

실험결과, 잡음환경에서 39유사음소를 이용한 경우가 48유사음소를 이용한 경우보다 평균적으로 17%와 28% 향상된 단어인식률과 문장인식률을 얻을 수 있었다. 이는 39유사음소 단위가 한국어 연속음성인식에 더 적합하고, 잡음환경하의 연속문장인식에서도 더 효과적임을 확인할 수 있다.

5. 결 론

본 논문은 잡음환경하에서의 한국어 연속음성인식에 효과적인 문맥의존 음향모델 수에 대한 연구로서 유사음소단위 수에 따른 인식 성능을 비교, 평가 하였다. 연속음성인식에 이용되는 문맥종속모델의 경우 변이음을 고려하여 모델이 작성되므로 이를 고려하면 기본 음소를 48음소로부터 39음소로 줄일 수 있다. 39음소의 인식에 대한 유효성을 확인하기 위하여 48음소와의 인식성능 비교 평가를 수행하였다. 또한, 실제 잡음환경에서도 유효한 문맥의존 음향모델을 생성하기 위해서 잡음이 부가된 음성데이터베이스를 구성하였으며, 이 음성 데이터베이스를 각 음소별 HM-Net음향모델에 학습시켜 연속음성인식 실험을 수행하였다.

실험결과, 무잡음 환경하의 연속음성인식에서 단어인식률은 약 7%, 문장인식률은 약 17%의 인식성능향상을 보였다. 잡음환경하의 연속음성인식에서도 평균적으로 39유사음소단위를 기본음소로 사용하였을 경우, 단어 인식률은 약 17%, 문장인식률은 약 28%의 성능향상을 보였다. 따라서 39음소가 발음변이가 빈번히 일어나는 연속음성인식 환경에서 48음소보다 효과적인 음소구성임을 알 수 있었으며, 잡음환경하의 연속음성인식에서도 효과적임을 확인할 수 있었다.

참고 문헌

[1]서준배, 김주곤, 김민정, 정호열, 정현열, "강건한 한국어 연속음성인식을 위한 유사음소단위에 대한 연구,"한국음향학회 학술발표대회 논문집, 제23권 제1(s)호, pp. 37-40, 2004

[2]임영춘, 오세진, 김범국, 정현열, "HMnet을 이용한 한국어 음소인식에 관한 연구," 한국음향학회 영남지회 학술발표대회 논문집, 제7권, pp. 50-53, 2000.

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

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

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

[6]M. Suzuki, S. Makino, A. Ito, 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.

[7]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

[8]김유진, 김회린, 정재호, "인식 단위로서의 한국어 음절에 관한 연구," 한국음향학회지, 제16권 제3 호, pp. 64-72, 1997.

[9]박기상, 석수영, 정호열, 정현열, "잡음환경에서의 음성인식을 위한 특징파라미터 비교분석," 한국음향학회지, 제22권 1호, pp. 141-144, 2003.