음성 인식에서 훈련 및 인식 과정에 사용되는 대상 어휘의 차이에 대한 음향 모델의 성능 평가

Performance Evaluation of Acoustic Models According to Differences between Vocabularies in Training and Test Phases of Speech Recognition

김 회 린*, 이 항 섭*, 권 오 욱**
(Hoi Rin Kim*, Hang Seop Lee*, Oh Wook Kwon**)

※이 연구는 정보통신부의 지원으로 이루어진 결과물입니다.

요 약

본 논문에서는 ETRI에서 개발한 가변 어휘 음성 인식기의 어휘 독립 음향 모델링 방법을 기술하고, 이 모델의 어휘 종속, 어휘 독립 및 어휘적응 성능을 평가하기 위하여 다양한 고립단어 및 연속음성 DB에 대하여 실험한 결과를 분석하였다. 평가를 위하여 사용한 음성 DB로는 고립단어 음성으로 POW(Phonetically Optimized Words) 3848, PBW(Phonetically Balanced Words) 445, PBW 452, 호텔예약 244 단어, 게임 제어용 단어 등이며, 연속음성으로 일반 문장 음성 및 연속 숫자음을 이용하였다. 성능 분석 결과 40개 음소 모델만으로도 비교적 높은 인식률을 보여 주었지만, 어휘독립의 경우는 어휘 종속에 비하여 성능이 크게 낮았고, 특히 대상 어휘가 숫자음, 알파벳, 연속음 등의 경우에는 POW 데이터나 PBW 데이터만 가지고는 우수한 가변 어휘 음성 인식기를 구현하기에 한계가 있음을 알 수 있다. 또한, 훈련 데이터의 어휘와 평가 데이터의 어휘가 비슷할 경우에는 변이음 모델을 사용하면 음소 모델만을 사용할 경우에 비하여 그 성능이 우수하였지만, 일반적인 어휘독립의 상황에서는 효과가 별로 없음을 알 수 있었다.

ABSTRACT

In this paper, we describe the vocabulary-independent acoustic modeling method of the variable vocabulary speech recognizer developed in ETRI, and evaluate it in the vocabulary-dependent, independent, and adaptive situations with a variety of isolated and continuous speech databases. For the evaluation data, we use POW(Phonetically Optimized Words) 3848 DB, PBW(Phonetically Balanced Words) 445 DB, PBW 452 DB, Hotel Reservation 244 word DB, and PC game control DB as isolated word DB. As continuous speech DB, we use general sentence speech DB and continuous digit speech DB. Experimental results show relatively high performance with only 40 phoneme models, but the performance of vocabulary-independent case is much worse than that of vocabulary-dependent case as we expected, especially in digits, alphabets, and continuous speech. This means that it is difficult to train acoustic models robustly only with POW and PBW DB for the variable vocabulary speech recognizer including digits, alphabets, and continuous speech. In the case with similar vocabularies in training and evaluation data, allophone models perform better than phoneme models, but it is not in general vocabulary-independent situations.

I. 서 론

가변 어휘 음성 인식기[1,2,3]는 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 자동으로 교체하여 단어 모델들을 재구성 하므로 이론적으로 무제한의 임의의 단어를 주어진 단어 목록 내에서 인식할 수 있다. 이러한 가변 어휘 음성 인식기를 구현하기 위해 사용하는 최초의 훈련용 음성 데이터 베이스에 모든 한국어 음소가 포함되어 있지 않거나, 모두 포함하고 있더라도 음소 환경이 충분히 다양하지 않으면, 이를 이용하여 어휘에 무관한 음성 인식기를 만들 때 인식 성능의 저하를 피할 수 없게

* 한국전자통신연구원 음성신호처리팀
** 한국전자통신연구원 음성언어팀
접수일자 : 1998년 6월 2일

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.[4]

하지만 어휘 독립 음향 모델은 앞서 기술한 바와 같은 풍부한 음성 DB를 사용하여 훈련해도 어휘 종속 음향 모델에 비하여 인식 성능이 저하된다. 따라서 어휘 독립 음향 모델의 개선 방향을 결정하기 위해서는 다양한 음성 데이터에 대하여 그 성능의 차이를 구체적으로 비교 분석할 필요가 있다.

본 논문에서는 ETRI에서 개발한 가변 어휘 음성 인식기의 어휘 독립 음향 모델링 방법을 기술하고, 이 모델의 어휘 종속 및 어휘 독립 성능을 평가하기 위하여 다양한 고립단어 및 연속음성 DB에 대하여 실험한 결과를 분석한다. 평가를 위하여 사용한 음성 DB로는 고립단어 음성으로 POW 3848, PBW 445, PBW 452, 호텔예약 244 단어, 게임용 단어 등이며, 연속음성으로 일반 문장 음성 및 연속 숫자음을 이용하였다.

II. 어휘독립 음향 모델링

1. 개 요

가변 어휘 음성 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[5]을 사용하여 다수의 화자로부터 음성을 수집하여 사용하고, 음소의 다양성을 모델 구조에 반영하기 위하여 음소만이 아닌 변이음까지의 상세 모델링을 함으로써 각 모델의 정확도를 향상시켰다.

본 연구에서 사용한 변이음 정의는 음성학적 지식을 기반으로 한 규칙[6]을 기본으로 하되, 이를 일부 단순화하여 재정의 하였다. 이와 같이 정의된 변이음 추출법에 따라 생성될 수 있는 변이음의 종류는 이론적으로 2,551개가 가능하다. 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. 여기에서 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.

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.

2. 어휘독립 음소 및 변이음 모델링을 위한 훈련용 음성 DB

가변 어휘 음성 인식 시스템에 사용될 음소 및 변이음 모델의 첫단계 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 set씩 모두 8 set을 각종 모델의 훈련에 사용하며, 이 중에서 남성음 3 set과 여성음 2 set은 수작업으로 음소 경계가 labeling되어 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.

3. 특징 추출 및 훈련

특징 벡터 추출 과정은 다음과 같다. 먼저, 10 msec (160 samples) 마다 256 point FFT를 수행하여 21개 밴드의 특징 벡터를 구하고, 이로부터 13차 PLP (perceptually linear prediction) 특징 벡터를 구한다. 구해진 특징 벡터로부터 dynamic feature를 구하기 위해 현재 프레임의 전후 각 3 프레임을 이용하는 탭 수 7개의 FIR filter를 사용하여 first-order dynamic feature를 얻고, 이 두 가지 벡터를 연결한 26차 벡터에 음성입력에 사용되는 마이크 등의 채널 차이를 제거하기 위하여 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 code-

word를 가지는 phonetically-tied SCHMM으로 모델링 하며, 모델의 구조는 3-state left-to-right (no skip path) model로 정의되어 있다. 이때 묵음은 state 수 1개인 모델로 정의하였다. 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.[7]

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.

III. 고립단어 음성 인식 실험

1. 음성 DB

구현된 가변 어휘 음성 인식기의 성능을 평가하기 위하여 5가지 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 DB, PBW 452 DB[1], 호텔예약용 244 단어 음성 DB, 그리고 PC게임을 음성으로 제어하기 위한 게임 제어용 DB를 사용하였다. 표 1에 나타나 있는 바와 같이 각 데이터는 훈련과 평가를 위하여 다시 세분화 되어 있다. 특히 POW 데이터는 1차 훈련용 데이터와 2차 훈련용 데이터를 분리하여 사용하였으며, PBW 445 및 PBW 452 데이터는 어휘 종속에 대한 성능을 평가하기 위하여 2차 훈련에 참여할 데이터를 분리하였다. 호텔예약 데이터는 단지 어휘 독립 평가용으로만 사용하였고, 게임 제어 데이터는 마이크를 desktop용과 headset용으로 분리하여 수집하여 사용하였다. 또한 모든 평가용 DB의 녹음, 양자화 및 끝점 검출은 훈련용 DB와 거의 동일하다. 수집된 데이터의 잡음환경은 POW, PBW 445, PBW 452, 호텔예약 데이터의 경우 평균 신호 대 잡음비(SNR)가 약 30dB 정도이지만, 게임 제어 데이터는 주변 잡음이 비교적 많이 포함되어 있고, 각 화자의 발성 크기도 낮아서 평균 신호 대 잡음비가 약 17dB 정도로 낮은 데이터이다.

1) PBW 452 DB는 과학기술부가 지원하고 국어정보학회가 주관하는 "국어정보처리 기술 개발 사업"의 일환으로 원광대에서 수집한 보급용 음성 데이터베이스임.

2. 실험 결과 및 분석

표 2 및 표 3에 가변 어휘 음성 인식기의 화자독립, 고립단어 인식 성능이 기술되어 있다. 우선 표 2에서 Baseline 시스템은 음소 경계 정보를 포함하고 있는 POW 1차 set을 사용하여 각 음소 모델을 훈련한 것이다. 어휘 종속의 성능과 비교하기 위하여 2차 훈련에서는 PBW 445 및 PBW 452 데이터를 포함하여 훈련하였다. 이때는 초기 음소 모델로 Baseline 시스템에서 사용한 음소 모델을 사용하고, bootstrapping 훈련의 HMM 파라메터 재추정 단계(iteration)를 2회까지만 반복하였다. 단어 인식 시의 검색과정으로는 Viterbi beam search를 사용하였으며, 이때 beam threshold는 2.5로 하였다. 이 값은 여러 가지 다른 값에 대하여 실험해 본 결과이다.

먼저 POW 데이터를 사용한 어휘종속 단어 인식률은 71.2%로 저조하였는데 이것은 검색 사전이 3,848개의 어휘로 구성되어 있고, 이들이 1 음절에서 9 음절까지 골고루 분포하고 있어서 적은 수의 음절로 구성된 경우 오인식률을 증가 시켰으며, 음향 모델도 단지 40개의 음소 모델만을 사용했기 때문인 것으로 분석된다. 이 기본 시스템을 이용하여 PBW 445 및 PBW 452 데이터에 대한 어휘독립 성능은 비록 어휘수가 1/8 정도로 크게 줄었지만 비교적 좋은 성능을 보여준다. 특히 PBW 452 데이터에 대한 성능이 좋았는데 이것은 어휘들간의 상이성이 보다 컸기 때문이다. 호텔예약 데이터에 대한 성능은 어휘수가 가장 적었음에도 불구하고 기본 시스템의 성능이 66.2%로 가장 저조하였는데 이것은 이 DB의 어휘 내에 숫자음 및 영어 알파벳 단어가 약 2/3 정도나 되어 유사한 발성이 매우 많았기 때문이다.

2차 훈련에서 HMM 파라메터 재추정을 1회 수행한 음소 모델들에 대한 성능 평가에서는 어휘종속의 경우 POW 데이터에 대한 ERR(Error Reduction Rate)이 7.6%였고, PBW 445 데이터에 대해서는 25.3%, PBW 452 데이터에 대해서는 45.6%를 보여주었다. 어휘독립의 경우는 호텔예약 데이터에 대하여 3.6%를 얻었다. 이러한 결과를 분석하면, POW 데이터의 경우 훈련 데이터 내용이 어휘수 및 데이터 양의 증가로 인하여 약간의 개선이 이루어졌다. PBW 데이터들에 대해서는 큰 폭의 개선이 이루어졌는데 이것은 어휘적응의 결과이며, 특히 PBW 452 데이터에 대해서는 훈련 데이터의 양이 크게 반영되어 더욱 큰 개선이 이루어졌다. 또한 어휘독립 실험의 결과인 호텔예약 데이터의 경우에는 훈련 데이터의 양이 크게 증가했음에도 불구하고 성능 개선이 미미하였는데 이것은 호텔예약 데이터의 어휘 내용이 역시 어려운 타스크임을 보여준다. 결국 숫자음이나 알파벳과 같은 어휘를 처리하기 위해서는 이러한 데이터를 훈련에 적극 반영하는 것이 중요함을 보여준다.

음소 모델의 훈련에 있어서 iteration 수가 1에서 2로 증가할 때 일반적으로는 인식 성능이 개선되지만, 본 실험에서는 전반적으로 저하되었다. 그 이유는 다음과 같다. 본 인식 시스템의 훈련 과정은 크게 두 단계로 나뉘어 수행된다. 하나는 각 음소 모델의 코드북을 전체 음성 데

표 1. 고립단어 음성 DB 목록
Table 1. List of isolated word speech databases.

DB 종류	어휘수	전체 DB			훈련용 DB			평가용 DB		
		화자수	어휘당 반복횟수	총 단어수	화자수	어휘당 반복횟수	총 단어수	화자수	어휘당 반복횟수	총 단어수
POW 3848	3,848	남자 40명 여자 40명	10회	38,480	1차 : 남자 24명 여자 16명	5회	19,240	남자 8명 여자 8명	2회	7,696
					2차 : 남자 32명 여자 32명	8회	30,784			
PBW 445	445	남자 21명 여자 19명	40회	17,800	2차 : 남자 15명 여자 15명	30회	13,350	남자 6명 여자 4명	10회	4,450
PBW 452	452	남자 38명 여자 32명	70회	31,640	2차 : 남자 27명 여자 23명	50회	22,600	남자 11명 여자 9명	20회	9,040
호텔예약 244	244	남자 9명	9회	2,196	X	X	X	남자 9명	9회	2,196
게임 제어	32	남자 48명 여자 32명	Desktop 콘덴서 Mic.: 3회*80명 = 240회 (209회)	6,656	남자 38명 여자 26명	3회*64명 = 192회 (167회)	5,362	남자 10명 여자 6명	3회*16명 = 48회 (42회)	1,294
			Headset 콘덴서 Mic.: 3회*80명 = 240회 (149회)	4,780		3회*64명 = 192회 (119회)	3,824		3회*16명 = 48회 (30회)	956

이터의 Viterbi segmentation 결과에 의해 다시 생성하는 단계로서 여기서 iteration으로 표현한 과정이고, 다음 단계는 단지 forward-backward 알고리즘에 의하여 각 코드워드의 weight 값을 update하는 내부 iteration 과정을 수행하는 과정이다. 본 실험에서는 이 내부 iteration 과정을 각 5회 반복하였는데 그 결과 생성된 모델 파라메터가 부적절한 값으로 변화된 것으로 분석된다.

특정 응용분야에 대한 실험으로 PC 게임 제어용 32 어휘 데이터에 대한 실험이 표 3에 정리되어 있다. 기본 시스템에 대한 완전 어휘독립의 성능은 평균 86.0%로 저조하였는데 이것은 앞서 기술한 바와 같이 데이터의 SNR이 17dB 정도로 잡음 수준이 비교적 높았기 때문이다. 훈련에 게임 제어 음성을 반영한 어휘적응의 결과는 음소 모델의 경우에도 96.3%의 성능을 보여주었다. 이러한 어휘적응 과정에서는 환경에 대한 적응도 함께 이루어졌기 때문에 어휘독립에 비하여 크게 높은 성능을 보여주었다. 훈련 시 POW 데이터만을 이용한 기본시스템의 음소 모델을 기반으로 게임 제어용 음성만을 이용하여 재훈련한 경우 음소 모델의 성능은 어휘적응한 경우에 비하여 ERR이 45.9%였다. 한편 단어 모델을 사용할 경우 단어 당 동일한 수(50)의 코드워드를 정의할 경우 ERR이 21.6%, 단어별로 음소수에 비례하는 코드워드 수를 적용할 경우 ERR은 75.7%로 인식률이 99.1%에 이르렀다. 이러한 결과는 결국 특정 응용 분야에 따라 훈련에 사용할 음성 데이터의 선정 및 모델의 선정에 따라 성능에 큰 차이가 있음을 보여준다.

표 2. 고립단어 음성 인식 실험 결과 (I)
Table 2. Experimental results for isolated word speech recognition (I).

훈련용 DB / 모델 ＼ 평가용 DB	Baseline 시스템 • POW 1차 set • 음소 모델	• POW 2차 set + PBW 445 + PBW 452 • 음소 모델	
		Iteration = 1	Iteration = 2
POW 3848	71.2%	73.4%	69.6%
PBW 445	77.5%	83.2%	80.3%
PBW 452	84.2%	91.4%	90.7%
호텔예약 244	66.2%	67.4%	61.8%

표 3. 고립단어 음성 인식 실험 결과 (II)
Table 3. Experimental results for isolated word speech recognition (II).

훈련용 DB / 모델 ＼ 평가용 DB		Baseline 시스템 • POW 1차 set • 음소모델	• POW 2차 set + 게임 제어 • 음소모델		• 게임 제어 • 음소모델		• 게임 제어 • 단어모델			
							단어 당 50개 codeword 사용		단어 당 음소수 비례 codeword 사용	
			Iter. = 1	Iter. = 2	Iter. = 1	Iter. = 2	Iter. = 1	Iter. = 2	Iter. = 1	Iter. = 2
게임 제어	Desktop	85.3	96.6	91.7	97.9	97.9	97.1	97.0	99.1	98.6
	Headset	86.7	95.9	93.0	98.1	98.2	97.0	96.9	99.0	98.6
	평균	86.0	96.3	92.3	98.0	98.0	97.1	97.0	99.1	98.9

IV. 연속 음성 인식 실험

1. 음성 DB

연속 음성에 대한 가변 어휘 음성 인식기의 성능을 평가하기 위하여 표 4에서와 같이 2가지 종류의 DB를 사용하였다. 첫번째 데이터는 한 문장 당 평균 11.8개의 비교적 긴 어절 수로 이루어진 총 291개의 일반적인 문장 음성이고, 두번째 데이터는 /공/을 포함하는 11개의 숫자음으로 한 문장 당 3개에서 7개의 연속되는 숫자음으로 이루어져서 평균 5개의 숫자음으로 구성되는 음성 데이터이다. 이 연속 숫자음 데이터는 훈련용으로도 사용하여 어휘종속에 대한 평가도 수행하였다.

표 4. 연속 음성 DB 목록
Table 4. List of continuous speech databases.

DB 종류	어휘 수	전체 DB			훈련용 DB			평가용 DB		
		화자수	총 문장수	총 단어수	화자수	총 문장수	총 단어수	화자수	총 문장수	총 단어수
문장 음성	2,327	남자 3명 여자 4명	291	3,428	X	X	X	남자 3명 여자 4명	291	3,428
연속 숫자음	11	남자 90명 여자 50명	5,169	약 2,6000	남자 68명 여자 37명	3,880	약 19,500	남자 22명 여자 13명	1,289	약 6,500

2. 실험 결과 및 분석

표 5에서 먼저 기본시스템에 대한 성능은 삽입 오류를 포함 했을 경우 45.8%, 삽입 오류를 제외 했을 경우 48.1%의 단어 인식률을 보여주었다. 저조한 성능의 원인은 단어 단위의 훈련 데이터 사용, 어휘독립 실험 환경, 그리고 40개 음소 모델만을 사용한 때문이다. 훈련 데이터에 PBW 445와 PBW 452 데이터를 포함한 경우에는 비록 기본 시스템에서와 동일한 성능 저하 요인을 가지고 있음에도 불구하고 성능이 58.0%로 개선되어 ERR이 22.5% 이었다. 이 결과는 앞서 호텔예약 단어 인식 실험에서 ERR이 저조했던 것과는 큰 대조를 보여주고 있다. 이 결과로 볼 때, POW 데이터는 단어 인식 시스템에서는 가변 어휘 인식기를 위한 어휘독립용 훈련 데이터로써 적절하지만, 연속 음성 인식 시스템의 어휘독립용 훈련 데이터로는 보완할 여지가 있음을 증명한다.

표 6에서는 연속 숫자음에 대한 가변 어휘 음성 인식기의 성능을 보여준다. 이 표에서의 단어인식률은 삽입 오류를 포함한 결과이다. 먼저 기본시스템에 대한 성능은 단어인식률이 71.1%, 문장인식률이 22.5%로 매우 저조하였다. 이것은 앞서 호텔예약 단어 인식의 경우와 마찬가지로 숫자음에 대한 어휘독립 인식 시스템의 한계를 보여준다. 이를 보다 확실히 보여주기 위하여 POW 2차 set에 대하여 변이음 모델을 구하여 평가해 보면 성능이 더욱 저하되어 단어인식률이 61.8%, 문장인식률이 9.9%로 된다. 이것은 변이음 모델링으로 인한 각 모델의 정확도 개선의 효과 보다는 숫자음에서 요구되는 각 모델의 훈련 특성이 더욱 나빠진 결과임을 보여준다. 연속 숫자음만을 사용한 훈련을 통하여 얻은 음소 및 변이음 모델의 어휘 종속에 대한 실험 결과를 보면, 성능이 크게 개선되었고, 특히 변이음 모델의 경우가 역시 음소 모델의 경우 보다 우수함을 보여준다. 또한, 단어 모델을 사용하면 그 성능이 더욱 개선됨을 쉽게 알 수 있다.

표 5. 연속 음성 인식 실험 결과 (I)
Table 5. Experimental results for continuous speech recognition (I).

훈련용 DB / 모델 ＼ 평가용 DB	Baseline 시스템 • POW 1차 set • 음소 모델	• POW 2차 set + PBW 445 + PBW 452 • 음소 모델	
		Iteration = 1	Iteration = 2
문장 음성 - bigram 사용 - 단어인식률(%)	45.8 (48.1)	58.0 (60.4)	57.2 (59.8)

(주) 괄호 안의 성능은 삽입 오류를 제외 했을 경우임.

V. 결 론

본 논문에서는 ETRI에서 개발한 가변 어휘 음성 인식기의 어휘 독립 음향 모델링 방법을 기술하고, 이 모델의 어휘 종속 및 어휘 독립 성능을 평가하기 위하여 다양한 고립단어 및 연속음성 DB에 대하여 실험한 결과를 분석하였다. 평가를 위하여 사용한 음성 DB로는 고립단어 음성으로 POW 3848, PBW 445, PBW 452, 호텔예약 244단어, 게임 제어용 단어 등이며, 연속음성으로 일반 문장 음성 및 연속 숫자음을 이용하였다. 성능 분석 결과 40개

표 6. 연속 음성 인식 실험 결과 (II)
Table 6. Experimental results for continuous speech recognition (II).

훈련용 DB / 모델 ＼ 평가용 DB		Baseline 시스템 • POW 1차 set • 음소모델	• POW 2차 set • 변이음 모델		• 연속 숫자음 • 음소 모델 (묵음포함 14개)			• 연속 숫자음 • 변이음 모델 (묵음포함 25개)			• 연속 숫자음 • 단어 모델 (묵음포함 12개)					
											단어 당 50개 codeword 사용			단어 당 음소수 비례 codeword 사용		
			Iter. = 1	Iter. = 2	Iter. = 0	Iter. = 1	Iter. = 2	Iter. = 0	Iter. = 1	Iter. = 2	Iter. = 0	Iter. = 1	Iter. = 2	Iter. = 0	Iter. = 1	Iter. = 2
연속 숫자음 (no-gram)	단어인식률(%)	71.1	61.8	55.8	89.6	90.2	89.8	90.1	91.2	90.6	90.9	91.5	91.5	92.3	92.6	92.7
	문장인식률(%)	22.5	9.9	6.6	58.7	60.6	58.7	59.8	63.5	61.0	61.8	64.2	64.6	67.6	68.7	68.4

음소 모델만으로도 비교적 높은 인식률을 보여 주었지만, 어휘독립의 경우는 어휘종속에 비하여 성능이 크게 낮았고, 특히 대상 어휘가 숫자음, 알파벳, 연속음 등의 경우에는 POW 데이터나 PBW 데이터만 가지고는 우수한 가변 어휘 음성 인식기를 구현하기에 한계가 있음을 알 수 있다. 또한, 훈련 데이터의 어휘와 평가 데이터의 어휘가 비슷할 경우에는 변이음 모델을 사용할 경우 음소 모델만을 사용할 경우에 비하여 그 성능이 우수하였지만, 일반적인 어휘독립의 상황에서는 효과가 별로 없음을 알 수 있다.

이러한 결과를 종합해 볼 때, 가변 어휘 음성 인식기의 성능을 보다 개선시키기 위해서는 이 인식기를 적용할 대상 타스크 도메인에 따라 적절한 어휘적응 기능을 포함해야 하며, 적절한 수의 변이음 종류를 정의하여 훈련하는 것이 필요함을 결론 지을 수 있다. 어휘적응 방법으로 쉽게 생각할 수 있는 것은 기존의 화자적응 기능을 이용하는 것도 한가지 방법이 될 수 있다.[8]

참 고 문 헌

1. 김회린, 이황섭, "POW 3848 단어 인식기 구현 및 어휘 독립 실험," *제13회 음성통신 및 신호처리 워크샵(KSCSP'96) 논문집*, 13권, 1호, pp.127-130, 1996.
2. 이황섭, 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.
3. 김회린, 이황섭, "음성학적 지식 기반 변이음 모델을 이용한 가변 어휘 단어 인식기," *한국음향학회지*, 제16권, 제2호, pp.31-35, 1997.
4. Hoi-Rin Kim, "HMnet evaluation for phonetic environment variations of training data in speech recognition," *Jour. of ASK*, Vol.15, No.4E, pp.28-36, 1996.
5. Yeonja Lim and Youngjik Lee, "Implementation of the POW (Phonetically Optimized Words) algorithm for speech database," Proc. of ICASSP, pp.89-91, 1995.
6. 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.
7. M. Hwang and X. Huang, "Subphonetic modeling with Markov states - SENONE," *Proc. of ICASSP*, pp.I-33-36, 1992.
8. Oh-Wook Kwon, Chong-Kwan Un, Hoi-Rin Kim, "Performance of vocabulary-independent speech recognizers with speaker adaptation," *Jour. of ASK*, Vol.16, No.1E, pp.57-63, 1997.

▲김 회 린(Hoi-Rin Kim)
한국음향학회지 제16권 제2호 참조
한국전자통신연구원 음성신호처리팀 선임연구원

▲이 황 섭(Hang-Seop Lee)
한국음향학회지 제16권 제2호 참조
한국전자통신연구원 음성신호처리팀 선임연구원

▲권 오 욱(Oh-Wook Kwon)
한국음향학회지 제16권 제1E호 참조
한국전자통신연구원 음성언어팀 선임연구원