최대 사후 추정 화자 적응을 이용한 가변어휘 고립단어 음성인식기의 사무실 환경에서의 성능 평가

Performance Evaluation of Variable-Vocabulary Isolated Word Speech Recognizers with Maximum a Posteriori (MAP) Estimation-Based Speaker Adaptation in an Office Environment

권 오 욱*
(Oh Wook Kwon*)

※이 연구는 정보통신부의 지원으로 이루어졌습니다.

요 약

본 논문에서는 임의의 단어를 인식하기 위하여 음성학적으로 최적화된(phonetically-optimized word) 음성 데이타베이스를 사용하여 훈련된 가변어휘 고립단어 음성인식기의 실제 인식기 사용 환경에서의 성능을 평가하였다. 이를 위하여, 훈련 데이타베이스에서와 상이한 환경에서 수집된 음성학적으로 균형 잡힌(phonetically-balanced word) 고립 단어 음성을 테스트 데이터로 사용하였다. 테스트 데이타는 일반적인 사무실에서 작동하는 노트북 PC에서 내장 마이크를 사용하여 녹음되었다. 이렇게 녹음된 음성을 사용하여 고립단어 인식기의 인식률을 측정하였다. 이 인식기는 최대 사후(maximum a posteriori) 추정 알고리듬을 사용하여 화자의 변화에 적응하였다. 컴퓨터 모의실험 결과에 의하면 화자 적응을 하지 않은 기본 시스템은 깨끗한 음성에 대하여 81.3%에서 사무실 환경 음성에 대하여 69.8%로 인식률이 저하되었다. 사무실 환경 음성에 대하여, 비교사 점진(unsupervised incremental) 모드에서 최대 사후 추정 화자 적응 알고리듬을 적용하였을 경우에는 화자적응을 하지 않은 경우에 비하여 9%의 에러를 감소시키며, 50단어의 적응 단어를 사용하여 교사 묶음(supervised batch) 모드에서 최대 사후 추정 화자 적응 알고리듬을 적용하였을 경우에는 16%의 에러를 감소시켰다.

ABSTRACT

We evaluate performance of isolated word recognizers in an office environment. The recognizer is trained by a phonetically-optimized word (POW) speech database and hence it can recognize vocabulary of any tasks. We use maximum a posteriori (MAP) estimation-based speaker adaptation algorithm to cope with changes of speakers. To evaluate the recognizer, we use a phonetically-balanced words (PBW) as a test data recorded using a notebook PC with an internal microphone in an office environment. Simulation results show that the recognition accuracy of the recognizer without speaker adaptation in an office environment degrades 69.8% while the recognizer for clean speech is 81.3%. The recognizer in an office environment reduces 9% of recognition errors when MAP estimation-based speaker adaptation algorithm is applied in an unsupervised incremental mode, and reduces 16% of recognition errors in a supervised batch adaptation mode.

I. 서 론

본 논문에서는 개인용 컴퓨터(PC: personal computer)용 음성인식기의 실용화를 위하여 개발된 고립단어 인식기의 성능을 평가한다. 개발된 인식기는 임의의 단어를 인식할 수 있도록 훈련되어 있으며 인식 대상 어휘를 동적으로 바꿀 수 있으며, 환경이나 화자의 변화에 적응하기 위하여 화자 적응 기능을 가진다. 이 인식기는 인식 대상 어휘가 수시로 바뀔 필요성이 있는 음성 구동 웹브라우저와 같은 용도에 적합하다.

음성인식기의 성능향상을 위하여 화자적응이 널리 사용되고 있다. 이는 화자종속 인식기의 성능이 화자독립 인식기의 성능을 능가하며, 화자독립 시스템의 파라미터

*한국전자통신연구원 음성언어연구실
접수일자 : 1997년 11월 10일

를 화자의 특성에 맞도록 적응하는 화자적응 인식기의 성능은 화자종속 인식기의 성능에 근접해지기 때문이다. 반연속 은닉마코프모델(HMM : hidden Markov model)을 이용한 음성인식에서는 화자 적응에서 인식기 파라미터를 추정하는 방법에 따라서 최대사후(MAP : maximum a posteriori)추정 알고리듬, 변환을 이용하는 알고리듬, smoothing을 이용하는 알고리듬이 주로 연구되어 왔다.

MAP추정을 이용한 알고리듬[1-3]에서는 주어진 관측 샘플이 주어졌을때 파라미터의 사후확률을 최대화하도록 파라미터를 바꾼다. 선험밀도(prior density)가 가우시안일때, 최대사후추정 알고리듬은 선험파라미터와 최우도(maximum likelihood) 추정에 의하여 계산된 파라미터의 가중합으로 표시된다. 이때 가중치는 관측된 샘플의 갯수와 파라미터의 선험밀도의 분산에 의하여 결정된다. 적응 데이타가 충분할 경우에는 MAP추정을 이용한 화자적응 인식기는 화자종속 인식기에 수렴하게 된다.

변환을 이용한 알고리듬에서는 affine변환을 이용하는 것[4]과 선형회귀[5]를 이용하는 것이 있다. 두가지 모두 관측 샘플의 확률을 최대화하는 변환을 찾아내어 이를 파라미터에 적용한다. 이것은 적응 데이타가 적을 경우에 우수한 성능을 나타내며, 적응 데이타가 충분한 경우에는 MAP추정 알고리듬보다 성능이 좋지 않은 것으로 알려져 있다. 최근에는 이 두가지를 결합하는 방법이 연구되고 있다.

Smoothing을 이용하는 알고리듬[6,7]에서는 파라미터 추정이 두 단계로 이루어진다. 첫 단계에서는 최우도 추정으로 얻어진 파라미터와 이에 대응하는 선험파라미터의 차이를 구한다. 그 다음, 그 차이값은 벡터필드이론[6]이나 마코프 랜덤필드 이론[7]을 이용하여 인접한 차이값으로 smoothing된다. 이러한 smoothing기법은 퍼지 smoothing기법과 유사하며, smoothing의 정도를 주의깊게 결정하여야 한다.

위의 세가지 화자 적응 알고리듬은 정해진 적응 문장을 사용자가 발성하도록 한 다음 인식기의 파라미터를 일시에 바꾸는 교사 묶음(supervised batch) 모드와, 매 발성마다 인식 결과를 이용하여 인식기 파라미터를 바꾸어 가는 비교사 점진(unsupervised incremental) [13] 모드에서 동작할 수 있다. 교사 묶음 적응의 경우 사용자는 많은 양의 적응 문장을 발성하여야 하므로 불편을 초래한다. 비교사 점진 적응은 별도의 적응 단계가 불필요하므로 사용자 편이성은 좋으나 잘못된 인식 결과에 대해서도 적응을 하는 경우가 발생하므로 인식률 측면에서는 불리하다.

본 연구에서는 구현이 비교적 용이하고 적응 데이타가 충분할 경우 화자종속시시템에 수렴하는 MAP 추정 알고리듬을 사용하였으며, 비교사 점진과 교사 묶음의 두가지 모드에 대하여 인식성능을 평가하였다. 인식기 훈련에 사용된 인식 어휘와 겹치지 않는 새로운 어휘를 대상으로 하여, 잡음이 첨가되지 않은 깨끗한 음성과 실제 사용환경인 사무실환경에서 녹음된 음성에 대하여 인식실험을 수행하였다. 다음 절부터 MAP 추정 화자적응 알고리듬과 컴퓨터 모의실험 결과를 기술한다.

Ⅱ. 최대사후추정을 이용한 화자적응 알고리듬

본 연구에서 사용된 음향모델링은 senone[8, 9]을 기반으로 하고 있으며, 특징벡터 O_t를 상태 j에서 관측할 확률밀도함수는 다음과 같이 표시된다.

$$b_j(O_t)=\sum_{k=1}^{K} W_{jk}\, N(O_t;\mu_k,\sum_k).$$

여기서 W_{jk}는 가우시안 성분에대한 가중치로서 분포(distribution)의 구성요소가 된다. $N(O_t;\mu_k,\sum_k)$는 평균벡터 μ_k, 공분산 행렬 $\sum_k$인 가우시안 확률밀도함수(pdf)이다. 위식의 가우시안 확률밀도함수(pdf)들은 상태 j에 할당된 코드북에 속한 것들만 사용된다. 1차원의 특징벡터를 가정할 경우 특징 벡터열 O를 관찰 하였을때 사후확률 $\Pr(\mu,\sigma^2|O)$을 최대화하도록 W_{jk}', μ', σ'^2를 추정한다[1-3, 10,15]. 이렇게 구해진 추정평균값과 원래의 평균값의 차이는 다음과 같이 표시된다.

$$\begin{aligned}\Delta\mu &\equiv \mu'-\mu_0\\ &=\frac{\sigma_0^2}{\frac{\sigma^2}{n}+\sigma_0^2}(\bar{\mu}-\mu_0)\\ &\approx\frac{n}{n+\alpha}(\bar{\mu}-\mu_0)\end{aligned}$$

여기서 n은 관측된 샘플수이며 μ_0는 선험적인 평균값이며, μ'는 추정된 평균값이며, $\bar{\mu}$는 샘플 평균값이며, σ_0^2는 선험적인 평균값의 분산이며, σ^2는 샘플로부터 구한 분산이다. 즉 α는 샘플로부터 구한 분산과 선험적인 평균값의 분산의 비라고 생각할 수 있다.

분포는 MAP추정에 의하여 다음과 같이 주어진다[3].

$$W_{jk}'=\frac{n_{jk}+\beta_{jk}}{\sum_{k=1}^{K}(n_{jk}+\beta_{jk})}$$

여기서 n_{jk}는 코드북 j, 코드워드 k에 속한 샘플수이다.

$$\begin{aligned}n_{jk}&=\sum_t \Pr(O_t\in codeword\ of\ codebook\ j|O_t)\\ &=\frac{W_{jk}N(O_t;\mu_k,\sum_k)}{\sum_{k=1}^{K}W_{jk}N(O_t;\mu_k,\sum_k)}\Pr(O_t\in state\ j)\end{aligned}$$

여기서 β_{jk}는 분포의 선험값이며 다음과 같이 근사화하여 사용하였다.

$$\beta_{jk}=\beta W_{jk}.$$

여기서 β는 상태에 독립적인 상수값이다. 이와 같이 MAP화자적응에 사용되는 선험파라미터를 α와 β로 간략화 함으로써 선험파라미터값을 추정하기 위한 절차가 불필요하다.

Ⅲ. 실험환경

3.1 기본시스템

인식기는 끝점검출, 특징추출, 탐색과정으로 구성된다. 끝점 검출은 에너지와 영교차율을 이용하는 방법을 사용하였다. 입력된 16 kHz, 16비트 음성신호는 매10 ms마다 256개의 샘플값을 하나의 블럭으로 하여 13차 perceptually linear prediction (PLP)계수[12]로 변환된다. 최종적으로 13차 PLP계수와 그 선형결합으로 구해지는 13차 미분치를 합하여 26차 특징벡터가 얻어진다.

인식기에 사용된 음소모델은 묵음을 포함하여 40개이며, 묵음은 1개의 상태로 그외 음소는 3개의 상태를 갖는 skip이 없는 left-to-right HMM으로 모델링되었다. 각 상태마다 다른 코드북을 사용하여 총118개의 코드북을 구성하였으며 각 코드북의 코드워드 갯수는 모두 50이었다.

훈련데이타로는 핸드셋 마이크를 사용하여 방음실에서 녹음된 8세트의 음성학적으로 최적화된 단어(POW: phonetically-optimized word) 3848개로 이루어진 음성데이타베이스[11]를 사용하였다. 화자수는 남 32, 여 32명이었으며, 각 화자는 481개의 단어를 발성하였다. 인식기의 파라미터는 수작업으로 구한 레이블 정보를 이용하여 훈련하였다. 2 세트의 POW단어를 사용하여 실험한 음소모델 단어인식기의 인식률은 71.4%이었다[14]. 인식기의 자세한 규격 및 성능은 [14]를 참고바란다.

3.2 인식실험 환경

본 연구에서는 실제상황에서의 음성데이타베이스 수집의 어려움을 피하고자 기존 음성데이타베이스를 최대한 이용하였다. 인식기의 성능을 평가하기 위하여 잡음이 없는 깨끗한 음성, 다른 워스테이션에서 음성을 스피커로 재생하여 PC에서 녹음한 음성, 깨끗한 음성에 잡음만을 PC에서 녹음하여 첨가한 음성에 대하여 실험하였다. 세번째 경우는 두번째 경우와 비교하여 워스테이션 스피커와 PC 사운드카드의 특성이 음성에 반영되지 않는다는 점이 다르다. 두번째 경우가 실제상황과 더 유사하지만 음성데이타를 다시 녹음하는 노력이 더 필요하며, 워스테이션 스피커의 특성에 의하여 음성이 왜곡된다는 점이 실제로 화자가 PC에서 음성을 발성한 경우와 다르다. 본 연구에서 사용한 테스트용 데이타로는 445개의 음성학적으로 균형 잡힌 단어(PBW: phonetically-balanced word)를 2회 발성한 고립단어 데이타베이스 PBW445중에서 1회분 발성을 사용하였다. 이 데이타베이스는 헤드셋 마이크를 사용하여 방음실에서 녹음되었다. 테스트에 사용한 화자는 40명 중에서 임의로 선택한 10명(남6, 여4)이었다.

본 실험에서는 노트북 PC의 내장마이크를 사용하여 16 kHz, 16비트로 샘플링한 음성신호를 인식하였다. 녹음된 음성신호는 일반 사무실환경에서 발생하는 하드디스크 냉각팬 소리, 발자국 소리, 문 여닫는 소리, PC에서 조금 떨어지는 곳에서 대화하는 소리, 전화벨 소리 등의 잡음을 모두 포함하고 있다. 본 연구에서 표시한 인식률은 끝점검출 결과를 포함하는 인식기 전체의 성능이다. 즉, 본 논문에서 평가하고자 하는 인식기는 훈련과 테스트 데이타 사이에 인식어휘, 화자, 잡음환경이 모두 다르다는 점이 기존의 일반적인 단어인식기 인식률 평가와 다른 점이다.

Ⅳ. 실험결과 및 토의

4.1 깨끗한 음성에 대한 인식 성능

본 연구에서는 실제 환경에서의 성능을 평가하기 위한 것이나 비교 기준으로 사용하고자 먼저 깨끗한 음성에 대하여 알고리듬의 성능을 평가하였다. 10명의 화자에 대한 화자독립 인식기의 인식률은 표 1에서와 같이 평균 81.3%로 나타났으며, 화자에 따른 편차가 커서 최저 72%, 최고 93%를 나타내었다. 여기서 445단어에 대한 인식률이 일반적인 인식기 성능보다 낮게 나타난 것은 훈련에 사용된 음성데이타베이스의 어휘와 실제 인식에 사용된 어휘가 다르기 때문이다.

4.1.1 비교사 점진 화자적응 인식 성능

비교사 점진 모드는 적응데이타가 별도로 제공되지 않으며 매 발성이 끝날 때마다 인식결과가 맞다고 가정하고서 이를 사용하여 HMM파라미터를 적응하는 모드이다. 이 경우 화자독립인식기의 인식률이 너무 낮으면 잘못된 적응데이타를 사용하여 적응하는 것과 같이 동작함으로 인식률이 향상되지 않는다. 화자적응 예비실험에서 비교사 점진 화자적응의 경우 분포를 적응하는 것은 인식률 향상에 도움이 되지 않았기 때문에 코드북의 평균만을 적응하였다. 이 경우 선험파라미터와 최우도 추정으로 구한 파라미터의 가중치를 결정하는 α값을 적절히

표 1. 깨끗한 음성에 대한 화자독립 및 비교사 점진 화자적응 인식기의 화자별 단어 인식률(%)

화자	f_cst	f_jmh	f_lsh	f_sjl	m_asw	m_hgh	m_kmh	m_lks	m_pcb	m_yys	평균
독립	90	88	72	85	73	86	78	73	93	75	81.3
적응	93	86	86	87	79	87	84	76	94	82	85.4

표 2. 깨끗한 음성에 대한 Confidence 가중치를 사용한 화자적응 인식률(%)

화자	f_cst	f_jmh	f_lsh	f_sjl	m_asw	m_hgh	m_kmh	m_lks	m_pcb	m_yys	평균
인식률	91	85	85	88	79	87	87	75	95	60	83.2

결정하여야 한다. 본 실험에서는 예비실험을 통하여 인식률이 가장 높아지는 $\alpha=400$을 미리 구하여 사용하였다. 분포의 적응은 인식률 향상이 되지 않고 일부 화자에 대하여 오히려 인식률이 저하되어 사용하지 않았다. 표 1에서와 같이 인식률은 평균 85.4%, 최저 76% 최고 94%로 나타났으며, 일부 화자에 대해서는 약간의 인식률 저하도 발생하였다.

발화수 증가에 따른 인식률의 변화를 살펴보기 위하여 그림 1에 최근 50단어의 인식률의 변화를 나타내었다. 즉 수평좌표 200에서의 인식률은 테스트단어 151-200사이의 50단어에 대한 인식률을 나타낸다. 그래프를 살펴보면 거의 모든 구간에서 적응을 한 경우의 인식률이 높게 나타났다.

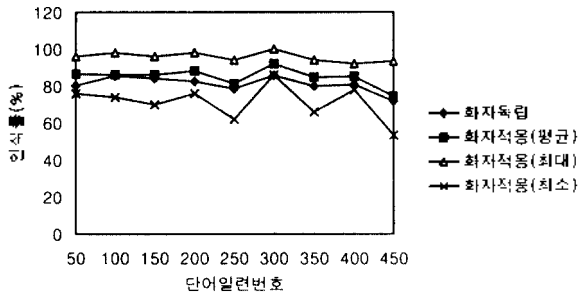

그림 1. 깨끗한 음성의 최근 50단어에 대한 화자독립 인식률 및 비교사 점진 화자적응 최소/최대/평균 인식률

비교사 점진 화자적응에서는 틀린 인식결과를 사용하여 적응하는 경우가 발생하기 때문에 인식결과를 조사하여 인식결과의 confidence에 따라서 그 발화에 의하여 적응되는 양을 조절할 수 있다. 인식결과에 대한 confidence, C를 사후확률의 개념을 이용하여 다음과 같이 정의하고 그 값이 분턱값이상이면 적응데이타로 사용하고 그렇지 않으면 적응데이타에서 제외하였다.

$$C=\frac{P(1)^{\gamma}}{\sum_{i=1}^{N}P(i)^{\gamma}}$$

여기서는 N-최적 인식결과 중에서 i-번째 관측확률이며, γ는 $P(1)$값의 기여도를 조절하는 상수이다. 분턱값을 0.8로하고 $N=10$, $\gamma=1$로 한 경우의 인식률을 표 2에 나타내었다. 인식률은 83.2%로서 confidence를 사용하지 않은 경우의 85.4%보다 낮게 나타났다. 이는 화자적응의 성능이 제대로 나타나려면 틀리기 쉬운 발화에 대하여 확실한 교사가 많이 존재하는 것이 가장 효과적인데, 이 방법에서는 인식결과의 신뢰도가 낮은, 따라서 인식이 잘 되지 않는 발성에 대해서는, 그냥 지나쳐 버림으로써 성능 향상이 기대에 못 미치는 것으로 판단된다. 새로운 방법이 요구된다.

4.1.2 교사 묶음 화자적응

교사 묶음 모드는 적응데이타와 그에 대응하는 transcription이 미리 주어지며, 모든 적응데이타의 입력이 완료된 다음에 HMM 파라미터를 일시에 적응하는 모드이다. 교사 묶음 화자적응에서의 분포적응은 코드북 적응보다는 작지만 어느 정도의 인식률 향상에 기여하지만, 그 크기가 작고 비교사 화자적응 실험결과와 같은 조건에서 비교하기 위하여 여기에서도 코드북의 평균만을 적응하였다. $\alpha=50$ 또는 100을 사용하고, 적응 단어의 갯수는 50과 100에 대하여 테스트하였다. 표 3과 같이 $\alpha=10$일때 적응단어 갯수를 50으로 할 경우 인식률은 84.6%, 100으로 한 경우 85.3%로 향상되었으며, $\alpha=50$일때 각각 84.9%와 85.4%로, $\alpha=100$일때 84.6%와 85.2%로 향상되었다. 이로부터 50개 정도의 적응데이타로도 어느 정도 인식률 향상을 보임을 알 수 있다. 이 경우 적응 데이타의 양도 중요하지만 인식기에 사용되는 음소모델별로 몇개의 샘플이 존재하느냐 하는 분제도 중요하다. 즉 모든 음소에 걸쳐서 존재할수록 적은 적응데이타로도 높은 성능 향상을 기대할 수 있다. 여기에서 사용한 적응데이타는

표 3. 깨끗한 음성에 대한 교사 묶음 화자적응 인식률(%)

	$\alpha=10$		$\alpha=50$		$\alpha=100$	
적응단어수	50	100	50	100	50	100
평균인식률	84.6	85.3	84.9	85.4	84.6	85.2

표 4. 사무실 환경에서 녹음된 음성에 대한 화자독립 및 비교사 점진 화자적응 인식기의 화자별 인식률(%)

화자	f_cst	f_jmh	f_lsh	f_sjl	m_asw	m_hgh	m_kmh	m_lks	m_pcb	m_yys	평균
독립	78	66	46	73	60	74	76	76	82	67	69.8
적응	83	75	43	75	61	80	80	76	87	66	72.6

갯수 50과 100일때 각각 한개 및 두개의 음소를 제외한 모든 음소를 포함하고 있다.

4.2 사무실 환경에서 녹음된 음성에 대한 인식 성능

웍스테이션 스피커를 통하여 재생하고 사무실 환경의 PC에서 녹음한 10명의 화자에 대한 화자독립 인식기의 인식률은 표 4에서와 같이 평균 69.8%로 나타났으며, 화자에 따라서 최저 46%, 최고 82%를 나타내었다. 이는 깨끗한 음성에 비하여 11.6%나 낮은 인식률이다.

4.2.1 비교사 점진 화자적응 인식 성능

$\alpha=400$을 사용하여 성능을 평가한 결과, 표 4에서와 같이 인식률은 평균 72.6%, 최저 43% 최고 87%로 나타났으며, 화자독립시스템에 비하여 9.0%의 에러감소율을 보였다. 화자적응 하지 않은 경우에 최저 인식률을 보인 화자의 인식률은 화자적응 후에 오히려 저하되었다. 이는 최대사후 추정을 이용한 화자적응에서는 화자독립 시스템의 인식률이 어느 정도 이상이 되어야 잘 동작함을 보여주고 있다. 이를 보완하기 위한 방법으로는 적응데이타의 양에 따라서 MAP적응에서 구해지는 추정평균과 원래평균의 차이값을 공유하는 방법도 고려할 수 있으나, 여기서는 이에 대한 실험은 생략하였다.

발화수 증가에 따른 인식률의 변화를 살펴보기 위하여 그림 2에 최근 50단어의 인식률의 변화를 나타내었다. 그래프를 살펴보면 대부분의 구간에서 적응을 한 경우의 인식률이 높게 나타나서 적응이 효과적임을 알 수 있다. 400-450구간에서의 낮은 인식률은 테스트 데이타로 사용한 음성데이타베이스의 단어 특성에 따라서 나타나는 것으로 판단된다.

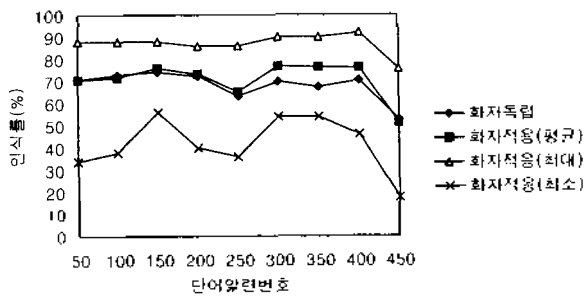

그림 2. 사무실 환경에서 녹음한 음성의 최근 50 단어에 대한 화자독립 인식률 및 비교사 점진 화자적응 최소/최대/평균 인식률

4.2.1 교사 묶음 화자적응 인식 성능

인식률 향상을 위하여 먼저 교사 묶음 화자적응 기법을 적용한 경우의 인식률은 표 5와 같다. $\alpha=10$일때 적응단어 갯수를 50으로 할 경우 인식률은 74.4%, 100으로 한 경우 75.0%로 향상되었으며, $\alpha=50$일때 74.9%와 75.1%로, $\alpha=100$일때 73.8%와 75.0%로 향상되었다. 이 결과로부터 인식률 변화의 추세는 깨끗한 음성을 사용한 경우와 유사함을 알 수 있다. 사무실 환경에서의 화자적응에 의한 에러감소율은 $\alpha=50$이고 적응단어 갯수 100일때 깨끗한 음성에서의 21.4%보다 감소하여 17.5%로 나타나서 화자적응의 효과가 약간 저하되었음을 알 수 있다.

표 5. 사무실 환경에서 녹음된 음성에 대한 교사 묶음 화자적응 인식률(%)

	$\alpha=10$		$\alpha=50$		$\alpha=100$	
적응단어수	50	100	50	100	50	100
평균인식률	74.4	75.0	74.9	75.1	73.8	75.0

4.3 잡음첨가로 얻은 음성데이타에 대한 인식 성능

PC에서 사무실 환경의 잡음만을 녹음한 다음 이 신호를 깨끗한 음성 신호에 첨가한 경우의 인식률을 조사하였다. 이 실험에서는 음성부분에 대해서는 훈련데이타베이스 수집시에 사용한 마이크 특성과 테스트 데이타에서 사용한 마이크 특성이 동일하다고 생각할 수 있으므로, 동일한 마이크를 사용하고 잡음환경 만 다른 경우의 인식률 변화를 볼 수 있다는 점에서 의미가 있다.

표 6의 인식결과를 살펴보면 화자독립시스템의 인식률은 평균 73.2%이며, 비교사 점진 화자적응을 할 경우에는 76.2%로서, 11.4%의 에러감소율을 나타내었다. 이는 마이크 채널 특성 변화와 잡음이 동시에 존재하는 앞의 경우보다 높은 인식률이다. 발화수 증가에 따른 인식률의 변화는 사무실 환경에서의 실험결과와 유사하였다.

V. 결 론

본 논문에서는 PC용 음성인식기의 인식성능을 평가하기 위하여 깨끗한 음성, 사무실 환경에서 녹음한 음성, 녹음된 잡음의 첨가로 얻어진 음성에 대하여 인식실험을 수행하였다. 평가에 사용된 인식기는 POW 음성데이타베이스를 사용하여 음소를 인식단위로 하는 반연속 HMM 모델을 가지며 임의의 단어를 인식하도록 훈련되었으며,

표 6. 잡음첨가로 얻은 음성데이타에 대한 인식률(%)

화자	f_cst	f_jmh	f_lsh	f_sjl	m_asw	m_hgh	m_kmh	m_lks	m_pcb	m_yys	평균
독립	77	71	49	72	68	78	81	77	86	73	73.2
적응	82	74	55	77	65	84	82	80	89	74	76.2

성능 향상을 위하여 최대사후 추정을 이용한 화자적응 기능을 가진다. 445개의 PBW로 구성된 음성데이타를 사용하여 내장마이크를 장착한 노트북PC에서 인식률을 측정하였다. 인식실험에 사용된 어휘는 훈련에 사용된 어휘와 다르다는 것이 기존에 주로 수행된 단어인식기 성능 평가와 다른 점이다. 컴퓨터 모의실험 결과 사무실환경에서 화자적응을 하지 않은 화자독립인식기는 69.8%의 단어인식률을 나타냈으며, 비교사 점진 모드에서 동작하는 최대 사후 추정 화자적응 알고리듬을 적용한 경우에는 72.6%, 50단어의 적응단어를 사용한 교사 묶음 모드에서는 74.9%로 향상된 단어인식률을 보였다.

앞으로 인식기의 성능향상을 위하여 잡음적응을 추가하거나, 화자적응과 거절 기능을 결합하는 방안도 연구할 필요가 있다. 또한 confidence를 이용한 비교사 점진 화자적응에서의 성능 향상 방안과, 적은 적응 데이타를 가지고도 높은 인식률 향상을 위하여 최우도 선형회귀(maximum likelihood linear regression)[5] 방법을 결합하거나 추정평균값과 원래의 평균값의 차이값을 공유하는 방안도 연구되어야 한다.

참 고 문 헌

1. C.-H. Lee, C.-H. Lin, and B.-H. Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models," *IEEE Trans. Signal Processing*, vol. 39, pp. 806-814, Apr. 1991.
2. J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech, Audio Processing*, vol. 2, pp. 291-298, Apr. 1994.
3. Q. Huo, C. Chan, and C.-H. Lee, "Bayesian adaptive learning of the parameters of hidden Markov model for speech recognition," *IEEE Trans. Speech, Audio Processing*, vol. 3, pp. 334-345, Sept. 1995.
4. V. V. Digalakis and L. G. Neumeyer, "Speaker adaptation using combined transformation and Bayesian methods," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 4, pp. 294-300, July 1996.
5. C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech, Language*, vol. 9, pp. 171-185, Apr. 1995.
6. H. Hattori and S. Sagayama, "Vector field smoothing principle for speaker adaptation," in *Proc. Int. Conf. Spoken Language Processing*, Alberta, Canada, pp. 381-384, Oct. 1992.
7. B. M. Shahshahani, "A Markov random field approach to Bayesian speaker adaptation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Atlanta, GA, pp. 697-700, May 1996.
8. M.-Y. Hwang and X. Huang, "Subphonetic modeling with Markov states-senone," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, San Francisco, CA, pp. I-33-I-36, Mar. 1992.
9. M.-Y. Hwang and X. Huang, "Shared-distribution hidden Markov models for speech recognition," *IEEE Trans. Speech, Audio Processing*, vol. 1, pp. 414-420, Oct. 1993.
10. M. H. DeGroot, Optimal statistical decisions. New York: McGraw-Hill, 1970.
11. Y. Lim and Y. Lee, "Implementation of the POW (phonetically optimized words) algorithm for speech database," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Detroit, MI, pp. 89-92, May 1995.
12. H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "RASTA-PLP speech analysis technique," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, San Francisco, CA, pp. I-121-I-124, Mar. 1992.
13. Q. Huo and C.-H. Lee, "On-line adaptive learning of the continuous density hidden Markov model based on approximate recursive Bayes estimate," *IEEE Trans. Speech, Audio Processing*, vol. 5, pp. 161-172, Mar. 1997.
14. 김회린, 이황섭, "음성학적 지식 기반 변이음 모델을 이용한 가변 어휘 단어 인식기," 한국음향학회지, vol. 16, pp. 31-35, 1997. 2.
15. O. W. Kwon, C. K. Un, and H. R. Kim, "Performance of vocabulary-independent speech recognizers with speaker adaptation," *J. Acoust. Soc. Korea*, vol. 16, no. 1E, pp. 57-63. Mar. 1997.

▲권 오 욱(Oh-Wook Kwon)

1986년 2월 : 서울대학교 전자공학과(학사)

1988년 2월 : 한국과학기술원 전기 및 전자공학과(석사)

1997년 2월 : 한국과학기술원 전기 및 전자공학과(박사)

1988년 3월~현재 : 한국전자통신연구원 선임연구원

※주관심분야 : 음성인식, 음성신호처리, 영상신호처리