최대사후확률 추정법을 이용한 단어인식기의 잡음환경적응화

Noisy Environmental Adaptation for Word Recognition System Using Maximum a Posteriori Estimation

이 정 훈*, 이 시 욱**, 정 현 열**
(Jung-Hoon Lee*, Shi-Wook Lee**, Hyun-Yeol Chung**)

※이 논문은 1995년도 한국학술진흥재단의 공모과제 연구비에 의하여 연구되었음

요 약

본 논문에서는 채널왜곡과 부가잡음에 강한 한국어 단어 인식기 구현을 위해 사후확률추정법에 의한 환경적응화법을 제안하고 이 방법의 인식성능 향상에 대한 유효성을 확인하였다. 이를 위해 1)채널왜곡이 발생한 경우, 2)부가잡음이 첨가된 경우, 3)채널왜곡과 부가잡음이 동시에 존재하는 각각의 경우에 대해서 제안한 환경적응화법을 이용하여 인식실험을 수행하였다. 이때 회귀계수, 지속시간 정보와 같은 부가정보의 환경적응화에 대한 유효성도 검토하였다. 100단어에 대한 환경독립, 화자독립 인식실험을 수행한 결과, 1)의 경우에 대해서는 9.0%, 2)의 경우에 대해서는 75%이상, 3)의 경우에 대해서는 11%~61.4%의 인식률 향상을 보여 사후확률추정법에 의한 환경적응화 방법이 채널왜곡 및 부가잡음이 동시에 존재하는 음성에 대하여서도 유효함을 알수 있었다. 그러나 지속시간 정보의 인식에 대한 기여는 찾아볼 수 없었다.

ABSTRACT

To achive a robust Korean word recognition system for both channel distortion and additive noise, maximum a posteriori estimation(MAP) adaptation is proposed and the effectiveness of environmental adaptation for improving recognition performance is investigated in this paper. To do this, recognition experiments using MAP adaptation are carried out for the three different speech; 1)channel distortion is introduced, 2)environmental noise is added, 3)both channel distortion and additive noise are presented. The effectiveness of additive feature parameters, such as regressive coefficients and durations, for environmental adaptation are also investigated.

From the speaker independent 100 words recognition tests, we had 9.0% of recognition improvement for the case 1), more than 75% for the case 2), and 11%~61.4% for the case 3) respectively, resulting that a MAP environmental adaptation is effective for both channel distorted and noise added speech recognition. But it turned out that duration information used as additive feature parameter did not played an important role in the tests.

Ⅰ. 서 론

음성인식은 화자독립(Speaker Independent), 대어휘(Large Vocabulary), 연속음성(Continuous speech)의 인식 및 이해를 최종 목표로 하고 있으나 불특정화자가 제한없이 자연스럽게 발성한 음성을 정도높게 인식하는 실시스템 구현을 위해서는 아직도 많은 연구가 필요하다. 실제로 이용 가능한 음성인식 시스템 구현을 위해서는 개인의 발성속도의 차에 의한 변동뿐만 아니라 전화기나 마이크의 종류, 인식기의 주변 잡음, 실내음향 특성 등의 주변 환경 변화에 대해서도 영향을 받지 않도록 설계되어야 한다.

그러나 현재까지의 연구는 발성화자, 어휘수, 발성내용, 발성방법, 발성환경 등에 대해 일정한 제한을 두어 인식시스템의 학습과 인식을 수행하는 경우가 대부분이다. 이 경우 비교적 양호한 인식결과를 기대할 수는 있지만 시스템의 사용환경이 달라질 경우 즉, 마이크가 달라지든지 주변 잡음이 존재하는 환경에서는 필연적으로 인식률의 저하를 초래한다[1]. 이를 해결하기 위해 음성인식 장치에 전처리 단계를 두는 방법으로 RASTA처리, CMN 처리 등이 보고되어 지고 있으나 부가잡음 및 채널왜곡

*대우통신 통신망 연구단
**영남대학교 전기전자공학부
접수일자:1997년 1월 17일

을 동시에 제거하는 데 문제가 있다든지, 계산량의 증가로 인해 실시간 시스템 구성에는 어려움이 따른다[2-10].

한편 음성인식에 있어서의 화자적응화는 불특정화자 모델을 표준모델로 미리 학습시켜 놓고 다른 환경에서 발성한 소량의 적응화용 음성만으로 추가적인 학습을 실시하여 다른 환경의 특정화자 모델의 특성에 가깝게 하고 있다[11-13]. 이와같은 방법은 주위 잡음의 부가, 마이크의 변동 등과 같은 부가잡음, 스펙트럴 잡음등을 동시에 해결하기 위한 방법으로도 고려 될 수 있다. 적응화 방법에는 여러 종류가 있으나, 최대사후확률추정법 (Maximum A Posteriori Probability)을 이용하면 소량의 적응화 음성으로 새로운 환경의 화자에 대해 적응화 학습이 가능하여 실시간 시스템 구현에 있어 학습량을 줄일 수 있다. 필자들은 지금까지 실제 환경에서 사용가능한 단어인식 시스템 구현을 위한 연구를 계속하고 있다. 그러나 인식기의 학습환경과 평가환경의 차이, 사용 마이크의 변경에 의한 채널왜곡의 영향으로 실제로 사용가능한 정도의 충분한 인식률을 얻지 못하고 있다[14, 15]. 이와 같은 요인을 제거하기 위해서 환경적응화를 수행하면 인식률 향상을 기대할 수 있으며, 특히 최대사후확률추정법을 이용하면 적응화가 중단되어도 그 시점까지 최적인 파라메터를 추정할 수 있고 필요시 추가적으로 적응화를 수행해 파라메터의 정밀도를 향상시킬 수 있다[16].

따라서, 본 연구에서는 인식시스템의 설계시와 상이한 마이크를 사용함으로 인해 발생하는 채널왜곡과 실제환경에서 부가되는 잡음의 영향에 의한 음성인식기의 인식률 저하를 방지하기 위해 환경적응화 학습을 도입하고 이때 환경적응화는 최대사후확률추정법을 이용한다. 이를 이용한 적응화 학습방법을 환경독립, 화자독립 인식실험에 적용하고 적응화학습 전후의 인식결과를 서로 비교 검토하여, 사용 마이크 및 발성환경이 상이한 입력 음성 데이터에 대한 적응화의 유효성을 확인한다. 또 인식시스템의 성능향상을 위하여 인식에 있어서의 정적파라미터로 이용되는 멜켑스트럼 계수 외에 스펙트럼의 시간방향 변화를 나타내는 동적 특징 파라미터, 지속시간정보등의 적응화 인식에 대한 유효성도 검토한다.

이때 인식의 기본단위는 CHMM(Continuous Hidden Markov Model) 음소모델로 하고 단어인식을 위해서는 탐색 시간 면에서 좋은 성능을 보이면서, 상대적으로 적은 기억장소를 이용하는 OPDP(One Pass Dynamic Programming)법을 이용한다.

Ⅱ. 음성의 분석

2.1 정적 특징 파라메터 추출

정적 특징 파라메타의 추출과정을 그림 1에 보인다. 음성데이터는 BPF(75Hz~8KHz)를 거쳐 20KHz의 샘플링율과 12비트로 양자화된 ETRI 611 단어 데이터베이스를 16KHz로 다운 샘플링 하여 사용한다. 음성데이터는 Pre-emphasis 필터를 통과시켜 고역 강조된 후 16ms(256points)의 길이의 해밍 윈도우를 사용하여 5ms(89points)씩 쉬프트 시키면서 분석된다. 이로부터 20차의 LPC 계수를 구하고 14차 LPC 켑스트럼 계수를 구한후, 10차의 LPC 멜켑스트럼을 구하여 정적 특징 파라메터로 한다. 표1에 음성 분석 조건에 대해 나타낸다.

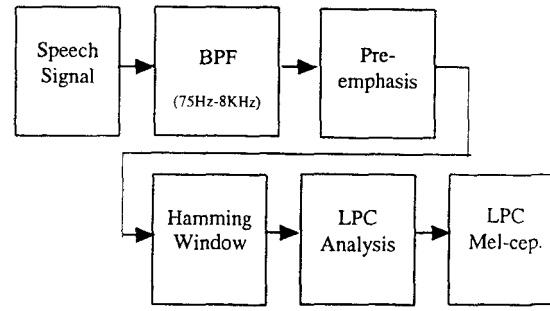


그림 1. 정적 특징 파라메타의 추출

표 1. 음성데이터의 분석조건

Speech Data	ETRI445	ETRI611
Sampling frequency	16kHz	20kHz→ 16kHz
Filtering	LPF 7kHz	BPF 75Hz-8kHz
Resolution	16bits	12bits
Hamming window	16ms (256points)	
Frame rate	5ms (80points)	
Aanlysis	14 order LPC analysis	
Feature parameters	10 mel-cepstrum coeff. 10 regressive coeff. Duration information	

2.2 동적 특징 파라메터 추출

본 연구에서는 음성신호 스펙트럼 내에서의 순시적인 변화를 나타내는 동적 특징 파라메터로서 회귀계수를 이용한다[2]. 회귀계수는 2.1절에서 구한 10차의 LPC 멜켑스트럼 계수로부터 각 프레임마다 추출하는데, 이때 회귀계수 $R_m(t)$는 시간 t를 중심으로 $2\delta+1$ 폭 만큼의 단위로 다음의 식(1)으로부터 구할 수 있다.

$$R_m(t)=\frac{\sum_{n=-\delta}^{\delta} nC_m(t+n)}{\sum_{n=-\delta}^{\delta} n^2} \quad (1)$$

여기서 $C_m(t)$는 발성 음의 t번째 프레임의 m번째 계수 값이고 $R_m(t)$은 여기에 해당하는 회귀계수의 값을 의미한다.

2.3 음소의 지속시간정보

음소 단위의 HMM을 작성할 경우에 길게 발성된 음소를 충분히 모델링하기 위해서는 자기천이를 가지는 left-to-right 모델로 가정하여 사용한다. 그러나 일반적으로 음소는 그 종류와 발성속도에 따라 지속시간 분포의 변동이 매우 심하다. 파열음이나 패쇄자음을 수반하는 음절의 중성모음은 그 길이가 매우 짧아서 수십 ms이내이나, 유성음이나 마찰음 등은 그 지속시간이 수백 ms이상되는 경우도 있다. 또한 대부분의 음소의 지속시간 분포는 종 모양의 분포를 가지는데 끝부분은 시작부분에 비해 서서히 감소한다. 그리고 어떠한 음소도 지수분포를 가지지는 않는다.

일반적인 음소를 기본단위로 HMM을 이용하여 지속시간의 분포를 모델링 할 경우 시간이 경과함에 따라 그 상태가 지속될 확률 $d(n)$는 식(2)와 같이 지수 함수적으로 감소하게 된다. 즉, 상태 i에 대해 이상태에 n시각(n프레임 입력) 머무를 확률은

$$d(n) = a_{ii}^{n-1}(1-a_{ii}) \tag{2}$$

로 나타나 n의 증가와 더불어 지수 함수적으로 감소한다. 그러나 실제음소의 지속시간은 감마나 포아송 분포를 가지므로 음성을 1차 마르코프모델로 모델링할 경우, 식(2)를 이용해서는 음소의 지속시간을 제대로 표현하지 못하는 단점이 있다. 이는 학습시 자기천이루프를 제거하고 지속시간분포 모델 (그림 2)을 도입하여 상태 i에서 τ시간 지속될 확률 $d_i(\tau)$를 구한 다음 이것을 새로운 파라메터로 추가함으로써 어느 정도 보완할 수 있다[12].

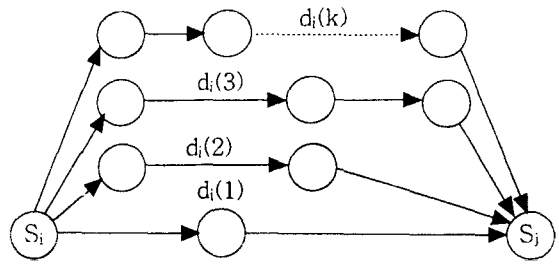


그림 2. 지속시간제어 모델

$$\text{단, } \sum_{\tau} d(\tau) = 1 \tag{3}$$

로 되는 구속조건을 만족하는 것으로 한다. 이 파라미터를 도입하면 연속출력분포 HMM의 경우, Baum-Welch 재추정 알고리즘은 다음과 같이 된다.

$$\alpha(i,j) = \sum_{j}\sum_{\tau \le T} \alpha(j, t-\tau)\,\alpha_{ji}\,d_i(\tau)\prod_{n=1}^{\tau} b_{ij}(o_{t+1-n}) \tag{4}$$

$$\beta(i,j) = \sum_{j}\sum_{\tau \le T-t} \alpha_{ij}\,d_j(\tau)\prod_{n=1}^{\tau} b_{ij}(o_{t+n})\beta(j, t+\tau) \tag{5}$$

여기서

$$\gamma_t(i,j,\tau) = \frac{\gamma(i, t-\tau)\,a_{ij}\,d_j(\tau)\prod_{n=1}^{\tau} b_{ij}(o_{t+1-n})\beta(j,t)}{P(o|M)} \tag{6}$$

이다.

Ⅲ. 환경 적응화 및 인식 알고리즘

3.1. 환경적응화 방법

특정화자나 불특정화자의 음성인식시스템을 최초 설계 환경과 다른 환경에서 미지의 화자가 사용할 경우에는 인식 대상 화자에 대해 소량의 음성 데이터로써 화자적응화를 수행해 인식시스템의 인식률을 높일 수 있는 화자적응화 방법이 많이 이용된다[13]. 그러나, 기존의 HMM을 사용한 음성인식기의 대부분은 ML(Maximum Likelihood)에 기반을 둔 Baum-Welch 학습법으로 파라메터를 재추정하고 있다. ML학습은 기본적으로 방대한 양의 학습 데이터와 각각의 모델이 서로 독립적이라는 가정을 기초로 한다. 그러나 실제 학습의 경우 각 모델들은 서로 독립적으로 보기 어렵고 학습 데이터 양도 상당히 제한되어 있어서 인식기의 변별력을 저하시키는 원인이 된다[5].

이 경우 최대사후확률추정법을 이용하면 적응화가 중단되어도 그 시점까지 최적인 파라메터를 추정할 수 있고 필요시 추가적으로 적응화를 수행해 파라메터의 정밀도를 향상시킬 수 있다. 따라서 본 연구에서는 최대사후확률추정법을 환경적응화 및 화자적응화에 도입하기로 한다.

식(7)은 최대 사후확률 추정식 나타낸 것이며 이 식을 이용하면 1개의 학습샘플이 주어진 경우에도 사후확률이 최대가 되도록 파라메터 Θ를 추정할 수 있다. 여기서 X_1, $X_2, \cdots, X_N$는 N개의 샘플을 나타낸다.

$$\max_{\Theta} P(\Theta|X_1,\cdots,X_N) = \max_{\Theta} \frac{P(X_N|X_1,\cdots,X_{N-1},\Theta)\,P(\Theta|X_1,\cdots,X_{N-1})}{\int P(X_N|X_1,\cdots,X_{N-1},\Theta)P(\Theta|X_1,\cdots,X_{N-1})\,d\Theta} \tag{7}$$

최대사후확률추정법을 이용해 정규분포의 평균과 분산을 동시에 추정하고자 하는 경우, 파라메터 Θ는 평균(μ)과 공분산(Σ)이 된다. 여기서, N개의 학습샘플을 이용한 평균벡터의 재추정식은 다음 식과 같이 나타낼 수 있다.

$$\hat{\mu}_N = \frac{\alpha\mu_o + \sum_{i=1}^{N} X_i}{\alpha + N} \tag{8}$$

또한, 공분산 행렬의 재추정식은

$$\bar{\Sigma}_N = \frac{1}{\beta+N}\{X_N X_N^T - (\alpha+N)\mu_N\mu_N^T + (\beta+N-1)\sum_{N-1} + (\alpha+N-1)\mu_{N-1}\mu_{N-1}^T\} \quad (9)$$

와 같다.

식(8)과 식(9)에서 α, β는 초기 HMM의 평균과 공분산 행렬을 추정하기 위해 사용한 샘플수 이며 수렴 속도와 관계가 있다. 따라서 이 값은 실험에 의해 재조정할 필요가 있지만 인식률에는 거의 영향을 끼치지 않는 것으로 알려져 있다[13]. 따라서 본 연구에서는 두 환경 및 모든 화자에 대해 적응화 계수 $\alpha=15$, $\beta=50$으로 일정하게 부여하였다.

3.2. OPDP법에 의한 인식

인식알고리즘은 탐색 시간 면에서 좋은 성능을 보이면서, 상대적으로 적은 기억장소를 이용하는 OPDP(One Pass Dynamic Programming)법을 사용한다. OPDP법은 일반적으로 음절이나 단어 단위로 작성된 표준패턴을 사용하여 연속음성 인식에 사용되어진다.

본 논문에서는 이를 개량하여 유사음소단위로 표준패턴을 작성하고 단어사전과 구문제어를 통하여 인식을 수행한다. 단어의 누적거리 계산은 입력의 첫 번째 프레임과 각 유사음소단위 표준패턴과의 Viterbi Score를 프레임에 동기시켜 가면서 계산한 후 저장한다[16]. 그리고 이를 단어사전중의 각 단어의 마지막 프레임까지 확장해서 총 누적거리 값이 최소가 되는 단어의 인덱스를 인식결과로 출력하게 하였다. 그림 3에 단어 인식시스템의 전체 흐름도를 나타낸다.

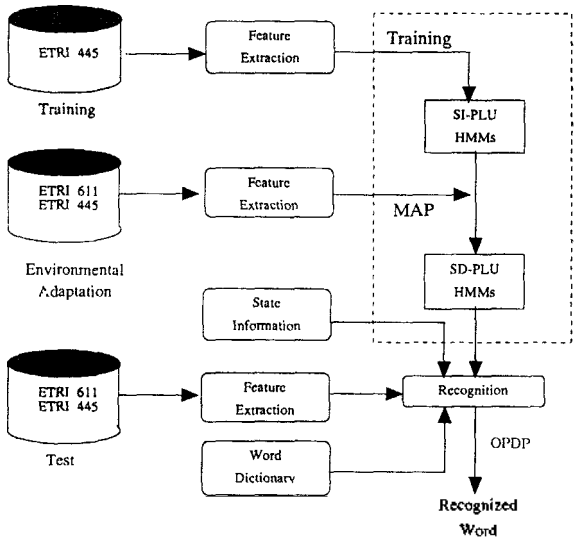


그림 3. 단어 인식 시스템의 흐름도

Ⅳ. 실험 및 고찰

4.1 음성 데이터 및 인식 모델

본 논문에서는 한국 전자통신 연구소(ETRI)에서 구축한 두 종류의 음성 데이터를 이용한다. 하나는 한국인 남성 22인의 2회 발성의 445단어(ETRI445)이며, 다른 하나는 같은 내용을 다른 남성 3인이 다른 마이크를 사용하여 각각 2회 발성한 611단어(ETRI611)이다. 표2에 음성 데이터 채록시에 사용한 마이크, 발성내용, 화자수를 나타낸다.

표 2. 음성데이터

Speech Data	ETRI445	ETRI611
Microphone	Dynamic Shure Headset	AKG Condensor Desktop
Content	Korean 445 Phoneme-balanced words	Korean 611 Phoneme-balanced words
Speaker	22 male persons	3 male persons

인식 실험에 있어서는 ETRI 445 단어 음성 데이터(ETRI 445)와 ETRI 611단어 음성 데이터(ETRI 611)를 사용한다. 초기 HMM 음소모델 작성을 위해서는 Dynamic Shure Headset 마이크를 사용한 ETRI 445 단어 중에서 10인의 남성 화자가 1회 발성한 100단어를 이용한다.

채널왜곡만 존재하는 경우, 부가잡음만 존재하는 경우, 채널왜곡 및 부가잡음이 동시에 존재하는 경우에 대해 환경적응화의 효과를 확인하기위해 ETRI 445 단어 음성 데이터 이외에 AKG Condensor (Desktop)마이크를 통하여 녹음되어진 ETRI611음성데이터 중 3인의 남성화자가 100단어를 1회 발성한 것을 추가로 사용하여 환경독립, 화자독립 및 적응화 인식실험을 수행한다.

초기 HMM 음소 모델의 구조는 4상태 1 혼합으로서 마지막 상태를 제외한 모든 상태에서 자기 천이를 가지는 순수한 left-to-right 모델로 한다. 이를 사용해 학습을 수행한 후, 자기천이를 제거하고 음소의 지속시간확률을 추가하여 학습시켜 인식모델로 한다. 그림4에 본 연구에서 사용한 연속분포 HMM모델의 구성을 보인다.

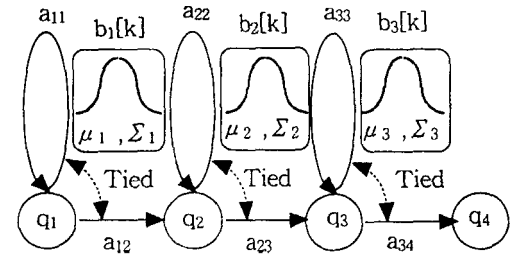


그림 4. 연속분포 HMM의 구성(4상태 1혼합)

4.2 인식 실험 및 고찰

인식실험에 있어서는 먼저 구성된 인식기의 기본 성능을 확인하기 위해 음성특징 파라메터를 점진적으로 부가해 가면서 100단어에 대한 화자종속 인식실험을 수행하였다. 이때 단어인식 알고리즘은 OPDP법을 사용하였다. 그 결과를 표 3에 나타내었다.

표 3. 특징 파라메타에 따른 화자종속 단어 인식

Feature parameterl	M	M+R	M+R+D
Average accuracy (%)	92.9	96.0	96.0

(M : Mel-cepstrum, R : Regression coeff., D : Duration)

표3의 결과로부터 초기 HMM 모델에 대한 화자종속 인식실험에 있어서는 특징 파라메터로 멜켑스트럼만을 특징 파라메터로 사용한 경우 92.9%의 인식률을 나타내었으며, 여기에 회귀계수의 특징을 부가하였을 경우 96.0%의 높은 인식률을 보였다. 또한 회귀계수와 지속시간의 특징을 함께 부가하였을 경우에도 96.0%의 인식률을 보임으로서 인식률의 차이가 없음을 알 수 있었다. 이는 음소를 인식의 기본 단위로한 100단어 정도의 화자종속 단어 인식에 있어서는 지속시간특징이 회귀계수의 특징에 포함되어 있는 것으로 해석된다.

이 결과를 바탕으로 시스템의 훈련시 사용한 마이크와 다른 특성을 가지는 마이크를 사용하였을 때 발생하는 채널왜곡의 인식에 대한 영향을 조사하기위해 Dynamic Shure Headset 마이크를 사용하여 녹음한 ETRI445 음성 데이터로 시스템의 초기 유사음소모델(PLU HMMs)을 작성, 훈련에 사용하고 평가시에는 AKG Condensor (Desktop)마이크로부터 채록한 ETRI611 음성 데이터를 이용하여 화자독립 인식실험을 수행하였다.

표 4. 채널왜곡을 가진 음성에 대한 환경적응화 효과 (단위:%)

Condition		Dynamic Shure Headset		AKG Condensor Desktop	
		Without Adaptation	With Adaptation	Without Adaptation	With Adaptation
Speaker	A	95	100	88	97
	B	96	100	78	96
	C	95	100	93	96
Average		95.3	100	87.3	96.3

인식결과 평균인식률은 87.3%로 나타나 동일한 마이크를 사용한 경우의 95.3%로에 비해 채널왜곡에 의해 8.0%의 인식률 저하가 있음을 알 수 있었다. 여기에 사후확률 추정법에 의한 환경적응화를 수행한 후에는 96.3%의 높은 인식률을 보여 9.0%의 인식률 향상을 가져와 이 알고리즘의 유효성을 확인할 수 있었다. 그 결과를 표 4에 나타낸다. 다음으로는 부가잡음에 의한 시스템의 성능열화를 확인하기 위해 기본모델 작성, 훈련 평가시에 동일한 마이크를 사용하고 평가용 데이터에 부가잡음을 첨가하여 인식실험을 수행하였다.

신호 대 잡음비(SNR : Signal-to-Noise Ratio)를 깨끗한 음성에 가까운 30dB에서부터 5dB단위로 10dB까지 부가잡음 레벨을 변화시키면서 화자독립 인식실험을 수행한 결과를 표 5에 나타내었다. 잡음이 심한 음성(SNR : 10dB 정도)에 대한 인식률은 환경적응화 수행전 평균 4.0%로 저조하였으나, 적응화 수행후에는 93.3%로 89.3%의 인식률 향상을 보였다. 그리고 신호 대 잡음비 20dB와 30dB 정도의 음성에 대해서도 환경적응화 수행전 인식률 6.7%와 24.6%가 환경적응화 수행후에는 96.3%와 99.7%로 나타나 각각 89.6%와 75.1%의 높은 인식률 향상을 보였다. 이로부터 잡음이 첨가된 음성에 대해서도 적응화 수행후에는 잡음정도에 상관없이 인식률이 75%이상 향상되어 사후확률추정법에 의한 환경적응화의 유효성을 확인할 수 있었다.

세번째로 마이크의 특성차이로 인한 채널왜곡 및 부가잡음이 동시에 존재하는 음성에 대하여 두 번째 실험과 동일한 방법으로 화자독립 인식실험을 수행하였다. 채널왜곡이 존재하고 부가잡음 정도가 심한 10dB에서는 적응화 수행전 9.3%의 평균인식률이 적응화 수행후에는 70.7%로 나타나 61.4%의 인식률 향상을 보였다. 그리고 20dB와 30dB정도의 잡음이 첨가된 음성에 대해서도 적응화 수행전 인식률이 각각 32.0%와 84.7%이었으나 환경적응화 수행후에는 76.7%와 95.7%로 나타나 각각의 경우에 대하여 44.7%와 11.0%의 인식률 향상을 보였다.

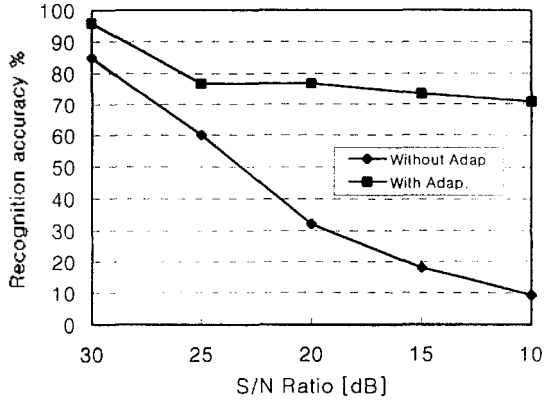
표 5. 부가잡음이 첨가된 음성의 환경적응화 효과 (단위:%)

S/N Ratio	30dB	25dB	20dB	15dB	10dB
Without adaptation	24.6	12.1	6.7	6.4	4.0
With adaptation	99.7	98.9	96.3	94.3	93.3
Increment	75.1	86.8	89.6	87.9	89.3

이 결과를 표 6에 나타내었다. 이상의 결과로부터 앞에서와 마찬가지로 사후확률추정법에 의한 환경적응화 방법이 채널왜곡 및 부가잡음이 동시에 존재하는 음성에 대하여서도 유효함을 알수 있었다. 여기에서 채널왜곡과 부가잡음이 함께 존재할 경우의 인식률이 부가잡음만이 첨가된 경우보다 더 높게 나타난 것은 양자화레벨의 차이에 의한 것으로 생각된다.

표 6. 채널왜곡과 부가잡음이 함께 있는 경우의 환경적응화 효과 (단위:%)

S/N Ratio	30dB	25dB	20dB	15dB	10dB
Without adaptation	84.7	60.0	32.0	18.0	9.3
With adaptation	95.7	76.7	76.7	73.3	70.7
Increment	11.0	16.7	44.7	55.3	61.4



V. 결 론

본 논문에서는 마이크의 채널왜곡과 실제 사용환경에서 부가되는 잡음에 강한 한국어 단어 인식기 구현을 위해 사후확률추정법에 의한 환경적응화법을 도입하고 이 방법의 인식성능향상에 대한 유효성을 확인하였다.

이를 위해 훈련과 평가시 두 개의 서로 다른 마이크를 사용하므로 인해 채널왜곡이 발생한 경우, 시스템의 사용환경 변화에 의한 부가잡음이 첨가된 경우, 채널왜곡과 부가잡음이 동시에 존재하는 각각의 경우에 대해서 제안한 사후확률추정법에 의한 환경적응화법을 적용하여 인식실험을 수행한 결과 다음을 확인할 수 있었다.

먼저 음성 인식기 자체의 성능을 평가하기 위해 음성특징 파라메터를 추가해 가면서 100단어에 대한 화자종속 단어인식실험을 수행한 결과 멜켑스트럼만 사용한 경우 92.9%, 회귀계수와 지속시간의 특징을 사용한 경우에는 96.0%의 인식률을 얻어 인식기의 성능이 비교적 우수함을 확인하였다. 그러나 회귀계수를 부가했을 경우와 회귀계수 및 지속시간 정보를 동시에 부가하였을 경우를 비교할 때 인식률의 차이가 없어 음소를 인식의 기본단위로 하는 100단어 정도의 화자종속 단어인식에서는 지속시간의 정보가 회귀계수 정보내에 포함됨을 알 수 있었다.

채널왜곡이 존재하는 경우에 대한 화자독립 인식실험 결과 평균인식률은 87.3%로 나타나 채널왜곡이 없는 경우의 95.3%로에 비해 8.0%의 인식률 저하가 있음을 알 수 있었다. 여기에 사후확률추정법에 의한 환경적응화를 수행한 후에는 9.0%의 인식률 향상을 가져와 이 방법의 유효성을 확인할 수 있었다.

다음으로는 부가잡음에 의한 시스템의 성능열화를 확인하기 위해 기본모델 작성, 훈련 평가시에 동일한 마이크를 사용하고 평가용 데이터에 부가잡음을 첨가하여 인식실험을 수행하였다. 신호 대 잡음비(SNR:Signal-to-Noise Ratio)를 30dB에서부터 5dB 단위로 10dB까지 부가잡음 레벨을 변화시키면서 화자독립 인식실험을 수행한 결과, 적응화 후에는 잡음정도에 상관없이 인식률이 75% 이상 향상되어 사후확률추정법에 의한 환경적응화의 유효성을 확인 할 수 있었다.

채널왜곡 및 부가잡음이 동시에 존재하는 음성에 대하여서는 11%~61.4%의 인식률 향상을 보여 사후확률추정법에 의한 환경적응화 방법이 채널왜곡 및 부가잡음이 동시에 존재하는 음성에 대하여서도 유효함을 알수 있었다.

이상의 결과로부터 본 논문에서 제안한 최대사후 확률추정법에의한 환경적응화법은 채널왜곡 또는 잡음이 부가되기 쉬운 음성인식기의 사용환경변화에 대해 유효하게 이용될 수 있음을 알 수 있었다.

*본 논문에서 사용한 단어 데이터베이스는 한국통신이 출연하여 한국전자통신연구소가 구축한 611단어음성 데이터베이스와 445 단어음성데이터베이스를 사용하였습니다.

참 고 문 헌

1. P.C.Woodland, M.J.F.Gales & D.Oye, "Improving environmental robustness in large vocabulary speech recognition", Proc. ICASSP'96, Vol. I, pp65-68, Atlanta.
2. J.E.Porter, S.F.Boll, "Optimal estimators for spectra restoration of noisy speech", Proc. ICASSP84, pp. 18A.2.1-4, 1984.
3. H.Hermansky, N.Morgan "RelAtive SpecTrAl (RASTA) processing in the speech analysis", Proc. 12th Speech Research Symposium, Rutgers University, June 1992.
4. H.Hermansky "Perceptual linear predictive (PLP) analysis for speech", J. Acoust. Soc. Am., pp. 1738-1752, 1990.
5. D.Van Compemolle, "Noise adaptation in a HMM speech recognition system", Computer, Speech and Language, Vol. 3, pp. 151-167, 1989.
6. 이우형, 정현열 "Robust 음성 인식기의 Front-End 설계를 위한 고찰", 한국음향학회 제11회 음성통신 및 신호처리 워크샵 논문집, 387-390.
7. 윤상호, 지상문, 오영환 "실험실환경 음성을 이용한 전화음성 인식에 관한 연구", 한국음향학회 제11회 음성통신 및 신호처리 워크샵 논문집, 391-394.
8. A.Acero and R.M.Stern, "Environmental robustness in the automatic speech recognition", Proc. ICASSP'90, pp. 849-852, 1990.

9. Alejandro Acero "Automatic and environmental robustness in automatic speech recognition", Kluwer Academic Publisher 1993.
10. H.Hermansky, N.Morgan, A.Bayya, P.kohn "Compensation for the effect of the communication channel in auditory-like analysis of speech", Proc. EUROSPEECH'91. pp. 1367-1370, Genova, 1991.
11. 한국전자통신연구소, "자동통역전화를 위한 요소기술개발(Ⅳ)", 1994, 12.
12. 中川聖一, "確率モデルによる音聲認識", 電子情報通信學會編, 1989.
13. 越川忠, "連續音聲認識システムにおけるHMMの話者適應化に關する研究", 修士學位論文, 1993.
14. 이우형, 정현열, "환경잡음에 강한 음성인식기의 FRONT-END", 제13회 음성통신 및 신호처리 워크샵 논문집, 356-360, 한국음향학회 (1996. 8).
15. 아시욱, 정현열, "음성인식 기능을 가진 주소입력검색 시스템", 제9회 신호처리 합동학술대회 논문집, 611-614, 한국음향학회 (1996. 10).
16. 이정훈, 정현열, "단어인식을 위한 환경적응화에 관한 연구", 제13회 음성통신 및 신호처리 워크샵 논문집, 50-54, 한국음향학회 (1996. 8).
17. Rabiner, Juang, "Fundamentals of Speech Recognition", Prentice-Hall International, Inc, 1993.
18. 甲斐充彦, "自然發話のための音聲認識システムに關する研究", 1996. 1.
19. Peter F.Brown, "Speech Recogniton By Statistical Methods", IBM 1985. 11.

▲이 정 훈(Jung Hoon Lee) 1971년 2월 15일생
1994년 2월 : 영남대학교 전자공학과 졸업
1997년 2월 : 영남대학교 대학원 전자공학과 졸업(공학석사)
1997년 1월~현재 : 대우통신 통신망 연구단
※주관심분야 : 음성분석 및 인식, 디지털 신호처리

▲이 시 욱(Shi Wook Lee) 1969년 11월 30일생

1995년 8월 : 영남대학교 전자공학과 졸업
1995년 8월~현재 : 영남대학교 대학원 전자공학과 석사과정
※주관심분야 : 음성분석 및 인식, 디지털 신호처리

▲정 현 열(Hyun Yeol Chung)
한국음향학회지 vol. 15, No. 3E, 1996. 참조.