멀티모달 인터페이스를 위한 음성 및 문자 공용 인식시스템의 구현

석수영[†] · 김민정[†] · 김광수[††] · 정호열[†††] · 정현열[††††]

요 약

본 논문에서는 음성과 온라인 문자를 단일시스템으로 인식할 수 있는 음성 문자 공용인식 시스템을 제안한다. 일반적으로 CHMM(Continuous Hidden Markov Model)은 음성인식과 온라인 문자인식을 위해 매우 유용한 도구로 잘 알려져 있으나, 인식을 위해서는 각각을 독립 시스템으로 구현하고 있어 추가적인 메모리와 계산량을 요구한다. 제안한 공용인식 시스템은 음성인식과 문자인식을 결합하기 위하여 이들을 동일한 CHMM 모델로 구성한 후 상태단위로 지속정보를 제어하는 OPDP(One Pass Dynamic Programming) 알고리즘을 통하여 음성과 문자를 인식할 수 있는 확률 통계적 시스템을 구현하였다. 음성은 MFCC(Mel Frequency Cepstrum Coefficient) 파라미터, 문자는 위치 변화량 파라미터와 비트맵 파라미터를 사용하였으며, MLE(Maximum Likelihood Estimation) 추정법을 이용하여 음소와 자소를 결합한 115개의 3상태 9천이 CHMM모델을 구성하였다. 공용인식기의 실험결과 음소 인식률 51.65%, 음성 단어 인식률 88.6%, 자소 인식률 85.3%, 필기체 단어 인식률 85.6%를 나타내어 공용인식의 유효함을 확인할 수 있었다.

An On-line Speech and Character Combined Recognition System for Multimodal Interfaces

Soo-Young Suk[†], Min-Jung Kim[†], Kwang-Su Kim[††], Ho-Youl Jung[†††] and Hyun-Yeol Chung[††††]

ABSTRACT

In this paper, we present SCCRS(Speech and Character Combined Recognition System) for speaker /writer independent, on-line multimodal interfaces. In general, it has been known that the CHMM(Continuous Hidden Markov Model) is very useful method for speech recognition and on-line character recognition, respectively. In the proposed method, the same CHMM is applied to both speech and character recognition, so as to construct a combined system. For such a purpose, 115 CHMM having 3 states and 9 transitions are constructed using MLE(Maximum Likelihood Estimation) algorithm. Different features are extracted for speech and character recognition: MFCC(Mel Frequency Cepstrum Coefficient) is used for speech in the preprocessing, while position parameter is utilized for cursive character. At recognition step, the proposed SCCRS employs OPDP (One Pass Dynamic Programming), so as to be a practical combined recognition system. Experimental results show that the recognition rates for voice phoneme, voice word, cursive character grapheme, and cursive character word are 51.65%, 88.6%, 85.3%, and 85.6%, respectively, when not using any language models. It demonstrates the efficiency of the proposed system.

Key words: speech, character, combined, recognition, SCCRS

이 논문은 산업자원부의 신기술실용화기술개발사업의 지원에 의해 이루어진 연구 결과물중 하나입니다.
접수일 : 2002년 4월 25일, 완료일 : 2002년 11월 19일
[†] 영남대학교 일반대학원 정보통신공학과
[††] 정회원, 경운대학교 컴퓨터전자정보공학부 전임강사
[†††] 정회원, 영남대학교 전자정보공학부 조교수
[††††] 정회원, 영남대학교 전자정보공학부 교수

1. 서 론

편리한 인간-컴퓨터 사이의 인터페이스를 구현하기 위해서는 음성, 제스처, 필기 등을 이용하여 표현된 사용자의 정보를 다방면으로 받아들일 수 있는 멀티모달 정보처리 시스템이 필요하다. 오늘날 멀티모달 시스템은 음성인식의 인식률 향상을 위해 입 모양 인식을 부가적으로 사용하는 시스템[1]과 화자인식을 부가적으로 사용하여 적합한 음향모델을 선택적으로 사용하는 시스템[2], 음성 인식과 손 제스처 인식을 통합한 시스템[3]등이 개발되고 있다. 그러나 이와 같은 시스템은 많은 계산량으로 인해 PC이상의 시스템을 요구하고 있다.

PDA(Personal Digital Assistant)와 같은 소형장치에서 멀티모달을 구현하기 위해 음성인식, 문자인식과 무선 통신을 이용하는 마이크로 소프트의 멀티모달 프로토타입인 MIPAD[4]가 발표되었다. MIPAD 시스템은 대어휘 음성인식을 위해 WINDOW CE환경의 PDA에서는 음성 전처리 과정만을 수행하여 무선랜으로 음성 파라미터를 전송하게 된다. 이후 서버에서 음성인식을 수행하여 PDA의 유저 인터페이스로 결과를 되돌려 주며, 이와 같은 시스템은 대어휘 연속음성을 소형 시스템에서 수행하기 위한 효과적인 구조이다. 그러나 PDA의 사용 환경이 항상 무선랜이 사용 가능한 장소에서 서버와 접속된 상태는 아니며, 또한 일반적인 음성명령어만을 위해 서버의 대어휘 음성인식엔진을 이용할 필요는 없다. 이를 위해 부동 소수점 연산 기능이 없는 저 성능의 시스템에서 음성인식을 추가하기 위한 연구가 진행되고 있다[5,6].

본 논문은 일반적인 PDA의 기본인터페이스인 온라인 문자인식에 추가적인 메모리양을 최소화하여 실시간 음성인식을 추가하기 위한 인식 시스템에 관한 것으로, 일반적인 음성인식과 문자인식은 독립적으로 수행되고 있어 대량의 메모리공간을 필요로 할 뿐 아니라 인식시간도 많이 소요되어 시스템 구현에 문제가 되고 있다. 이와 같은 문제점을 해결하기 위해서는 음성과 문자를 단일 인식기로 처리하는 방법이 생각될 수 있다.

현재 대부분의 음성인식기는 음성의 다양성을 수용하기 위해 HMM(Hidden Markov Model)을 이용한다[7] 또한, 온라인 필기체 문자인식의 경우에는 규칙화된 패턴 인식방법에서 HMM에 구조적, 전역적 지식을 적용하는 방법이 연구되어 좋은 성능을 나타내고 있다[8]. 따라서 본 연구에서는 음성인식과 문자인식에서 널리 사용되고 있는 HMM 인식 과정을 단일화한 공용 인식기를 구현하여 PDA와 같은 소형장치에서도 멀티모달리티를 구현할 수 있는 방법을 제안한다. 공용인식기를 구현하기 위해 음성인식 파라미터로 MFCC (Mel Frequency Cepstrum Coefficient)를 사용하고 문자인식을 위해 위치변화량 파라미터 및 비트맵 파라미터를 문자의 특징 파라미터로 사용하여 음성과 문자를 단일 CHMM(Continuous Hidden Markov Model)을 모델로 구성하여 인식을 수행한다. 또한, 출력 확률값 계산량을 감소시키기 위해 각 모델을 대각 공분산 행렬로 구성하고, 인식률 제고를 위해서 지속시간정보를 추가한 OPDP(One Pass Dynamic Programming) 탐색방법을 이용한다. 다음 2장에서는 음성과 문자인식을 통합한 공용인식 시스템에 대해 소개하고, 3장에서는 음성과 문자인식의 전처리 과정에 대해 소개한다. 4장에서는 제안한 3상태 9천이 CHMM모델과 지속시간 제어 OPDP 방법에 대해 소개하고, 5장에서는 음성과 문자의 각각의 음소 및 단어 인식실험과 결과에 대해 분석한 후 6장에서는 결론과 향후 연구방향에 대해 기술한다.

2. 공용인식 시스템

지능형 시스템은 음성인식, 문자인식, 필기/제스처 인식 등이 결합된 멀티모달 인식 시스템으로 발전하고 있다. 이를 휴대용 시스템에 적용하기 위해서는 기존에 각각 독립적으로 개발된 음성인식, 문자인식 시스템을 단일화함으로 시스템의 메모리 절감 및 보완성을 높일 수 있다. 이를 위하여 음성인식과 필기체 문자인식이 결합된 공용인식 시스템 개발이 필요하다. 그림 1은 음성과 문자를 입력받아 각각의 파라미터를 추출한 후 동일한 처리과정을 거쳐 결과를 추출할 수 있는 공용인식 시스템의 구성도를 나타내고 있다.

마이크를 통해 입력된 음성과 터치스크린으로부터 입력된 문자 데이터는 각각 다른 전처리 과정과 파라미터 추출 과정을 거쳐 동일한 파라미터 목적열로 만들어진다. 추출된 음성 및 문자 목적열은 레이블링 과정을 거쳐 통일한 CHMM모델로 학습되며, 음성은 유사음소단위 48개 모델과 문자는 67개의 자소 모델로 구성되어 총 115개의 1 가우시안 혼합수 CHMM모델

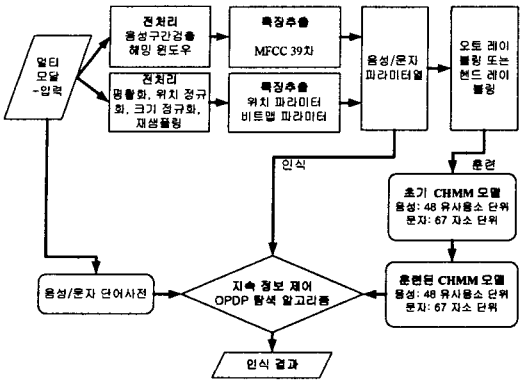


그림 1. 공용 인식 시스템 구성도

로 학습된다.

인식알고리즘은 음성/문자 동일하게 상태 단위로 지속 정보 제어가 가능한 OPDP 탐색방법을 사용하였다. 본 시스템은 모든 부분이 Visual C++ 환경 하에서 클래스단위로 구현되어 있으며, 전체적으로 CHMM 모델 작성을 위한 부분, 인식기부분으로 분리되어 구현되었다.

3. 전처리

전처리 과정은 잡음/잡영을 제거하고, 화자/필자의 개인적인 정보 표현을 정규화하여 특징을 추출하는 과정이다. 본 논문에서는 음성처리부의 경우 MFCC 파라미터를 이용하고 문자처리부는 위치 변화량 파라미터 및 비트맵 파라미터를 이용하였다.

온라인 문자 데이터는 타블렛 혹은 터치스크린으로부터 일정한 샘플링 간격으로 펜의 위치 정보를 나타내는 X, Y 좌표값이 표본화된다. 일반적인 문자인식의 전처리 과정은 크기 정규화, 위치 정규화(normalization) 과정, 잡영제거(noise reduction) 과정, 평활화 과정의 4단계로 구성된다[9]. 문자 데이터 입력시 필자에 따라 펜의 속도와 형태에 있어서 다양한 필체가 나타나게 되므로 인식기는 이와 같은 다양한 변화를 전처리과정에서 정규화할 필요가 있다. 정규화는 크기를 정규화하고, 실제 표본화될 영역으로 위치를 정규화하는 과정을 의미한다. 크기의 정규화는 글자단위의 외부분리가 이루어진 경우, x, y축 모두의 정규화를 수행할 수 있으나 본 시스템에서는 단어단위로 분리를 수행하므로 y축을 기준으로 이를 정규화하며, x축은 y축의 정규화 비율을 이용하여 크기가 수정된다. 위치

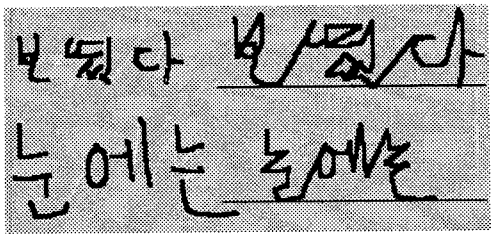


그림 2. 크기, 위치 정규화

정규화는 파라미터 추출시 좌표변화량만을 사용하지 않고, 절대좌표값을 이용함으로 위치정규화를 수행한다. 그림 2는 크기 위치 정규화후의 단어를 나타내고 있다.

평활화 과정은 일반적으로 필기 속도가 느리게 되면 손이 떨려서 글씨가 울퉁불퉁하게 표본화되는 문제로서, 이를 보정하는 방법으로 이웃한 점과 연관 지어 모든 점에서 평균화를 수행한다. 재추정된 $\hat{x}_{pj}$는 다음과 같은 식(1)에 의해서 구해진다.

$$\hat{x}_{pj} = C_{j-n} \cdot x_{pj-n} + \ldots + C_j \cdot x_{pj} + \ldots + C_{j+m} \cdot x_{pj+m} \quad (1)$$

여기서, n과 m은 선행과 후행의 점의 개수를 나타내며, $C_{j-n}, \ldots, C_{j+m}$은 각 점의 가중치를 나타낸다. 본 연구에서는 n=m=1을 사용하였다.

특징 파라미터 추출에 있어서는 전처리과정이 수행된 각 점으로부터 다양한 변형을 흡수 할 수 있는 특징을 사용하는 것이 유리하다. 이를 위해 본 시스템에서는 그림 3에서와 같이 터치스크린의 입력으로부터 사용자가 입력한 최소, 최대 좌표값을 기준으로 최소 좌표값을 "0"으로 이동하는 위치 정규화와 최대 좌표값을 일정한 값으로 정규화시키는 크기 정규화를 수행한다. 이후 식(2,3)에서와 같은 방법을 통해 재샘플링을 수행하여 모든 점을 새롭게 추출하게 된다.

$$px_i = \frac{sth*(x_j - px_{j-1})}{dis} + px_{j-1} \quad (2)$$

$$py_i = \frac{sth*(y_j - py_{j-1})}{dis} + py_{j-1} \quad (3)$$

$$dis = \sqrt{(px_{j-1} - x_j)^2 + (py_{j-1} - y_j)^2} \quad (4)$$

- sth: 샘플링 간격
- i: 재샘플링 과정 전의 열
- j: 재샘플링 과정으로 새롭게 만들어지는 열

전처리과정 수행 후 재샘플링(resampling)된 각 점

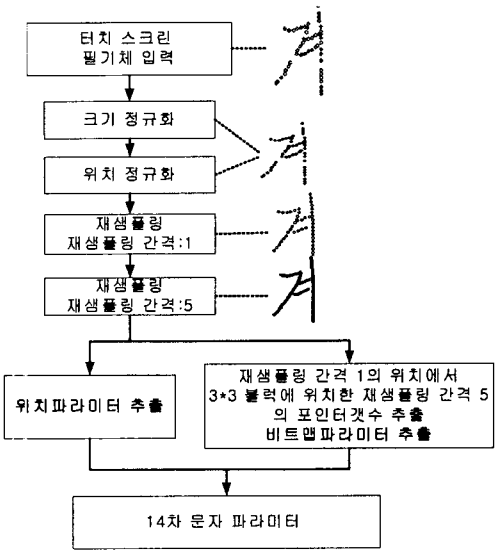


그림 3. 온라인 문자인식의 전처리 과정

으로부터 국부적인 점의 위치를 나타내는 y좌표 파라미터, 2차원의 국부적 각도 파라미터, 2차원의 국부적 만곡 파라미터에 전역적인 정보가 포함된 9차원의 비트맵 파라미터를 추가하여 총 14차의 파라미터를 추출한다[10].

음성인식의 전처리 과정은 크게 잡음제거(Noise Reduction), 음성끝점 검출(End Point Detection), 프리엠퍼시스, 헤밍 윈도우를 거쳐 로그에너지를 추출하고, 필터뱅크를 거쳐 MFCC 파라미터를 추출하게 된다. 이외의 파라미터로 에너지, Delta MFCC, Delta Delta MFCC를 추출한다. 그림 4는 음성처리부의 전처리과정을 나타내고 있다.

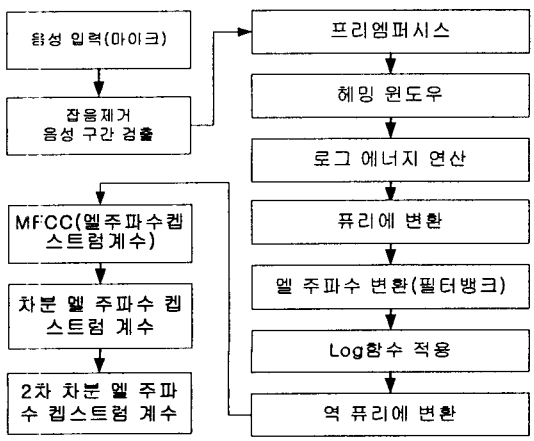


그림 4. 음성인식의 전처리 과정

실환경의 음성입력으로부터 잡음제거 방법은 스펙트럼상의 잡음성분을 계산하여 이를 차감하는 스펙트럼 차감법(Spectral Subtraction)을 이용한다. 스펙트럼 차감법은 수식 (5)~(7)과 같다.

$$S'(\omega) = [X(\omega) - \mu(\omega)]\exp(-j\theta_x(\omega)) \quad (5)$$

$$|S(\omega)| = |X(\omega)| - \mu(\omega)| \quad (6)$$

$$\mu(\omega)| = \frac{1}{M}\sum_{i=1}^{M}|N_{i(\omega)}| \quad (7)$$

여기서, $S(\omega)$는 음성신호의 스펙트럼, $X(\omega)$는 잡음에 의해 손상된 음성신호 스펙트럼, $N_i(\omega)$는 i프레임에서의 노이즈 스펙트럼을 나타낸다.

음성구간 검출 방법은 실환경의 음성구간 검출을 위해 비교적 작은 계산량으로 검출이 가능한 수정된 에너지와 영 교차율을 이용한 음성 구간 검출 방법을 사용하였다. 특히 임계값을 세분화와 검출된 구간의 음성/잡음 판단 기능을 부가하여 실시간 입력음성의 검출 성능을 향상시켰다[11].

프리엠퍼시스단계는 고주파수 성분을 강조하기 위해 수식 (8)을 이용하며, 여기에서 a값은 0.98을 이용하였다.

$$H(z) = 1 - \hat{a}z^{-1}, \quad 0.9 \le a \le 1.0 \quad (8)$$

윈도우 창 함수는 각 프레임의 시작과 끝에서의 신호의 불연속성을 최소화하기 위해 수식 (9)와 같은 헤밍윈도우를 이용하였으며, 또한 각 구간의 로그 에너지는 식(10)에서 의해 구해진다.

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1 \quad (9)$$

$$E_n = \left[\frac{1}{W}\sum_{i=1}^{W}|S_n(i)|\right] \quad (10)$$

문자 인식에 비해 음성인식이 어려운 이유 가운데 환경잡음을 제거하는 것이 온라인 필기의 잡영을 제거하는 과정에 비해 복잡한 면이 있으며, 또한 문자인식의 경우 전처리 과정 혹은 입력툴에 의한 대분류과정으로 인식기의 입력단위를 글자 단위로 분리 할 수 있으나, 음성의 경우 연속음성인식의 경우 문장, 고립단어 인식의 경우 단어 단위로 밖에 분리할 수 없는 어려운 면이 있다.

4. CHMM 통합모델

HMM은 음성인식 분야에서 현재 가장 널리 이용되

고 있는 통계적 모델로서 오늘날 시간열을 모델링 하는 분야에 널리 이용되고 있다. 화자간의 다양한 특성을 나타내는 음성인식과 필자간의 다양한 흘려 쓴 글씨가 나타나는 온라인 문자인식에 뛰어난 성능을 보이고 있다.

일반적으로 음성인식의 경우 문자인식에 비해 파라미터의 변화가 다양하게 나타난다. 따라서 시간열 상의 모델인 자기천이와 상태천이만을 가지는 Left-to-Right 모델로 구현시 별문제가 없으나, 문자인식의 경우 "ㄱ", "ㄴ"과 같은 자소는 기본적인 Baum-Welch 알고리즘을 이용한 MLE 학습시 2상태로 학습이 수행되며, "ㅣ"와 같은 자소는 1상태로 학습이 진행되는 현상이 발생한다. 따라서 NULL 천이가 없는 Left-to-Right 모델은 효과적이지 않으므로 그림 5와 같은 3상태 9천이 Left-to-Right 모델이 효과적이다.

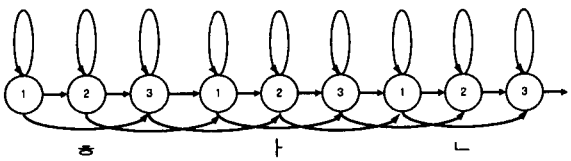


그림 5. 3 상태 9 천이 left-to-Right HMM 모델

4.1 Continuous HMM

제안한 공용인식 시스템은 3상태 9천이 3출력확률분포를 가지는 left-to-right CHMM 모델을 사용한다. 일반적으로, 각각의 혼합수는 전체 공분산, 혹은 대각 공분산 행렬로 가우시안 확률분포를 만들어 낸다. 식 (11)에서와 같이 전체 공분산 행렬을 사용하여 출력확률을 계산하는 경우 공분산의(주택청약예금) 행렬식 값과 역행렬의 계산이 필요하다.

$$b_j(y)=\frac{1}{\sqrt{(2\pi)^D \det U_j}}\exp\left\{-\frac{1}{2}(y-\mu_j)^T U_j^{-1}(y-\mu_j)\right\} \quad (11)$$

- U_j: 공분산 행렬

그러나, 식 (12)에서와 같이 대각 공분산 행렬을 사용하여 출력 확률을 계산하는 경우에는 행렬식계산과 역행렬을 위한 계산 과정이 필요하지 않게 된다. 또한 와 ξ_j 항으로 분리되어 항은 상태마다 계산하는 과정으로 줄어들고, 프레임마다의 계산은 ξ_j항만으로 줄어들게 된다.

$$b_j(y)=\frac{1}{\sqrt{(2\pi)^D \prod_{i=1}^{D}\sigma_{ji}^2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^{D}\frac{(y_i-\mu_{ji})^2}{\sigma_{ji}^2}\right\} \quad (12)$$

$$=S_j\cdot e^{-\xi_j(y)}$$

$$S_j=\frac{1}{\sqrt{(2\pi)^D \prod_{i=1}^{D}\sigma_{ji}^2}},\ \xi_j(y)=\frac{1}{2}\sum_{i=1}^{D}\frac{(y_i-\mu_{ji})^2}{\sigma_{ji}^2} \quad (13)$$

제안된 시스템은 계산량을 감소시키기 위해 출력확률 연산에 대각 공분산 행렬을 이용한다.

4.2 지속시간 제어 One-Pass DP

음성인식과 문자인식과 같이 시간열을 모델링 하는 HMM에서 지속시간정보의 이용은 인식률의 향상을 가져온다는 사실은 널리 알려져 있다[12]. 필기체의 경우 사람마다 글자를 쓰는 시간이 차이가 날수 있지만, 입력된 시간열상의 점을 재샘플링 하는 과정을 거치기 때문에 자소길이에 대한 프레임수가 나타나게 된다. 즉 "ㄴ"과 "ㄷ", "ㄹ"은 추출된 프레임의 수가 증가하게 됨으로써, 지속시간정보를 이용할 경우 인식률의 향상을 가져올 수 있다. 음성은 시간열을 정규화하는 전처리 하는 과정을 거칠 수 없기 때문에 문자에 비해 지속시간의 정보가 작아지게 된다. 본 연구에서는 지속시간을 제어하기 위해 지속 정보를 사용하는 방법[12]을 이용한다. 식 (14)는 상태 i에서 n 프레임동안 지속될 확률을 나타내고 있다.

$$d(n) = a_{ii}^{n-1}(a_{ii}-1) \quad (14)$$

이외에 상태 천이를 제어하기 위해 HMM모델 훈련시 각 인식 단위에서 나타나는 최소 프레임 수, 최대 프레임 수, 평균 프레임 수를 구해 프레임 단위로 가중치를 가하는 방법을 사용하였다. 가중치는 식(15,16)과 같다.

$$W(s)=(1+\frac{Avr_n/S_N-s_{duration}}{Avr_n/S_N}) \quad (15)$$

or

$$W(s)=2-\frac{S_N * s_{duration}}{Avr_n} \quad (16)$$

- Avr_n은 n 번째 HMM의 평균 프레임 수
- S_N은 자기천이를 가지고 있는 상태의 총수
- $s_{duration}$은 s 상태의 자기천이 프레임 수

5. 인식실험 및 고찰

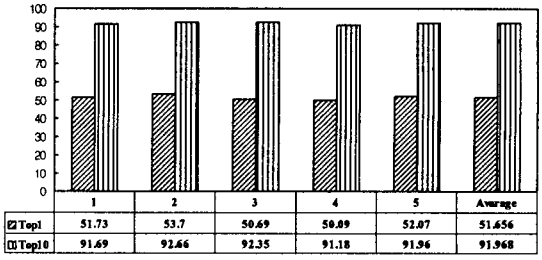
공용인식 시스템의 유효성을 확인하기 위해 음성에 대해서는 음소인식과 단어인식 실험을 수행하였으며, 문자에 대해서는 자소인식과, 단어인식 실험을 수행하였다. 화자 독립 모델을 구성하기 위한 음성 데이터는 한국전자통신연구원(ETRI)에서 작성한 PBW(Phoneme Balanced Words) 445단어 음성 데이터베이스 중 14인의 1회 발성을 이용하였으며, 필자 독립 모델을 구성하기 위한 문자 데이터는 KAIST에서 작성된 필기체 한글 DB중 10인의 1회 필기분을 이용하였다. 음성 및 문자 데이터의 분석 조건은 표 1과 같다.

음소 인식실험을 위한 평가용 데이터로 수작업으로 작성된 개별 음소를 이용하였으며, 10명의 16694개의 음소를 이용하여 화자 독립모델 구성 후 5명의 8354개의 음소를 이용하여 실험을 수행하였다. 표 2에서와 같이 실험결과 평균 음소인식률 51.65%를 나타내었으며, 후보 10위까지의 결과 96.61%를 나타내었다. 자소 인식실험을 위한 평가용 데이터로 수작업으로 작성된 개별 자소를 이용하였으며, 10명의 6080개의 자소로 필자 독립모델을 구성 후 5명의 3040개의 자소로 실험을 수행하였다. 표 3에서와 같이 실험 결과 평균 자소 인식률 85.36%를 나타내었으며, 후보 5위까지의 결과 96.01%의 인식률을 나타내었다. 자소인식률의 경우 음소와 달리 초성, 중성, 종성으로 구분이 가능함으로 "ㅏ"자소가 "ㄴ"자소로 오인식되는 결과를 배제한 결과이다.

표 1. 음성/문자 데이터 전처리 및 분석 조건

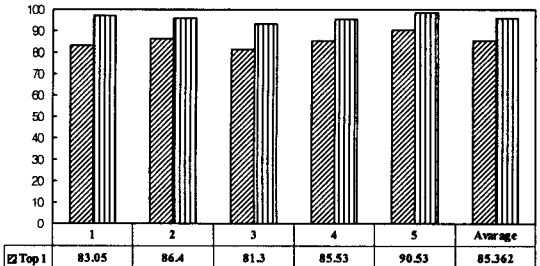
	음성	필기체 문자
전처리	16KHz 샘플링, 16bits 양자화, 헤밍 윈도우 5ms 쉬프트	100 샘플/초 평활화, 크기/위치 정규화, 재샘플링
특징 파라미터	MFCC 12 + Power 1 Delta (MFCC 12+ Power 1) Delta Delta (MFCC 12+Power 1)	Y축 위치 1, 방향 2, 만곡 2, 수정된 비트맵 9
데이터베이스	ETRI 단어 445 DB	KAIST 한국어 필기체 DB
모델	1 가우시안 혼합수 CHMM	

표 2. 음소 인식률



	1	2	3	4	5	Avarage
Top1	51.73	53.7	50.69	50.09	52.07	51.656
Top10	91.69	92.66	92.35	91.18	91.96	91.968

표 3. 자소 인식률



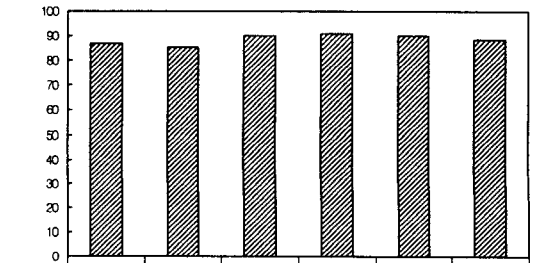
	1	2	3	4	5	Avarage
Top1	83.05	86.4	81.3	85.53	90.53	85.362
Top5	97.26	95.86	93.1	95.4	98.45	96.014

음성 단어 인식실험을 위한 단어 목록은 445개이며, 학습에 참가하지 않은 5인의 각 100단어를 대상으로 인식 실험을 수행하였다. 실험 결과를 표 4에 나타내었으며, 지속시간 제어 방법을 이용한 CHMM을 사용한 경우 화자독립 88.6%의 인식률을 나타내었다. 필기체 단어 인식실험을 위한 단어목록은 100개이며, 학습에 참가하지 않은 3인의 각 100단어 대상으로 인식 실험을 수행하였다. 실험 결과를 표 5에 나타내었으며, 지속시간 제어 방법을 이용한 CHMM을 사용한 경우 필자독립 85.6%의 인식률을 나타내었다.

6. 결 론

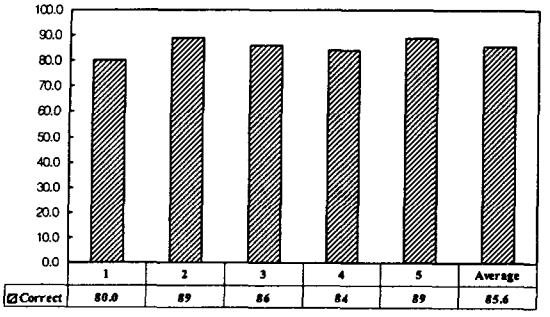
사용자가 보다 편리하고 융통성 있는 인간-컴퓨터

표 4. 음성 단어 인식률



	1	2	3	4	5	Avarage
Correct	86.74	85.39	90.11	90.78	89.89	88.582

표 5. 온라인 문자 인식률



사이의 인터페이스를 구현하기 위해서는 음성, 필기, 제스처 등을 통해 사용자의 의사를 충분히 표현, 전달하고 컴퓨터 측에서는 이들 정보를 인식, 이해하게 함으로써 얻고자하는 정보를 완전히 얻을 수 있는 멀티모달 정보처리 시스템 개발에 관한 연구가 필요하다. 그러나 현재의 멀티모달 시스템은 개별적인 시스템을 연결하여 구성된 시스템으로, 멀티모달을 구현하기 위해서 다중으로 메모리와 시스템의 계산량을 요구한다.

본 논문에서는 개별적인 인식과정을 수행하는 음성인식과 문자인식을 동일한 CHMM으로 모델을 구성한 후, 상태단위로 지속시간 제어를 추가한 OPDP 알고리즘으로 음성과 문자를 처리할 수 있는 음성/문자 공용 인식시스템을 제안하였다. 본 시스템은 전처리단과 특징 파라미터 추출과정은 음성과 문자처리부가 분리되어 독립적으로 수행되나, 모델 훈련, 인식 과정은 단일한 과정으로 이루어진다. 모델 구성시 자소들의 작은 위치 변화량을 보완하기 위해 NULL천이가 포함된 3상태 9천이 모델을 이용하였으며, 계산량을 감소시키기 위해 대각 공분산 행렬을 사용하였다. 음성의 특징 파라미터로 39차의 MFCC파라미터를 사용하고, 문자의 특징파라미터로 5차의 위치 변화량 파라미터와 9차의 비트맵 파라미터를 사용하였다. 공용인식기의 실험결과 음소 인식률 51.65%, 음성 단어 인식률 88.6%, 자소 인식률 85.3%, 필기체단어 인식률 85.6%를 나타내어 공용인식 시스템의 가능성을 확인하였으며, 이를 이용함으로써 PDA와 같은 소형의 머신에서도 음성인식과 문자 인식을 함께 사용 가능한 멀티모달리티의 구현이 용이하다.

참 고 문 헌

[1] Murai K, Nakamura S, "Real time face detection for multimodal speech recognition," *IEEE Multimedia and Expo 2002 Proceedings 2002*, Vol.2, pp.373-376, 2002.

[2] Chibelushi C.C, Deravi F, Mason J.S.D, "A review of speech-based bimodal recognition," *IEEE Multimedia*, Vol. 4, pp.23-37, March 2002.

[3] Lizhong Wu, Oviatt S.L, Cohen P.R, "From members to teams to committee-a robust approach to gestural and multimodal recognition," *IEEE Neural Networks*, Vol.13, pp.972-982, July 2002.

[4] Huang X, "MiPad: a multimodal interaction prototype," *IEEE Acoustics, Speech, and Signal Processing*, 2001. Vol.1, pp. 9-12, 2001.

[5] Delaney B, Jayant N, Hans M, Simunic T, Acquaviva A, "A low-power, fixed-point, front-end feature extraction for a distributed speech recognition system," *IEEE Acoustics, Speech, and Signal Processing 2002*, Vol.1, pp.793-796, 2002.

[6] Ning Bi, Garudadri H, Chienchung Chang, DeJaco A, Yingyong Qi, Malayath, N, Huang, W, "A robust speech recognition system embedded in CDMA cellular phone chipsets," *IEEE Acoustics, Speech, and Signal Processing 2002*, Vol.4 pp.3804-3807, 2002.

[7] X.D.Huang, Y.Ariki, M.A.Jack, *Hidden Markov Models for Speech Recognition*, Edinburgh Univ. Press, pp.167-185, 1990.

[8] 조성정, 김진형, "HMM기반 온라인 한글 인식에서의 구조적, 전역적 지식의 적용," 1998년 봄 정보과학회 학술발표논문집(B), pp.716-718, 1998.

[9] 성태진, 박승양, "문자 조합 규칙 학습에 의한 한글 온라인 필기 인식기의 설계," 한국정보과학회 추계 학술 발표 논문집, pp.223-226, October 1999.

[10] 석수영, 정현열, "CHMM 모델을 이용한 자소 분리 필기체 문자 인식," 한국음향학회 영남지회 (vol.7), pp.54-57, October 2000.

[11] 김춘영, 석수영, 정호열, 정현열, "에너지와 영교차율을 이용한 온라인 실시간 음성구간 검출," 한국음향학회 영남지회, October 2002.

[12] S.E.Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition," *Computer Speech and Language*, Vol.1, pp.29-45, March 1986.

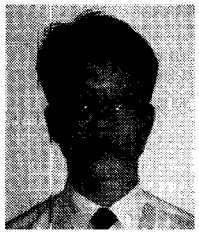
[13] M. K. Ravishankar, "Efficient algorithm speech recognition," Ph.D. thesis, *Computer Science Department*, Carnegie Mellon University, May 1996.

석 수 영

1998년 계명대학교 물리학과(이학사)
2000년 영남대학교 일반대학원 멀티미디어 통신공학과(공학석사)
2002년 3월~현재 영남대학교 일반대학원 정보통신공학과(박사수료)
관심분야 : 디지털신호처리, 음성인식, 문자인식

김 민 정

1999년 영남대학교 일반대학원 멀티미디어 통신공학과(공학석사)
2001년~현재 영남대학교 일반대학원 정보통신공학과(박사수료)
관심분야 : 디지털신호처리, 음성처리, 음성인식, 화자 인식



김 광 수

1994년 경남대학교 전자공학과(공학사)
1998년 영남대학교 일반대학원 전자공학과(공학석사)
2002년 영남대학교 일반대학원 전자공학과(공학박사)
2001년 3월~현재 경운대학교 컴퓨터전자정보공학부 전임강사
관심분야 : 음성분석 및 인식, 음성 및 오디오 신호처리, 음질평가

정 호 열

1988년 2월 아주대학교 전자공학과(공학사)
1990년 2월 아주대학교 전자공학과(공학석사)
1993년 2월 아주대학교 전자공학과(박사수료)
1998년 (프)리옹국립응용과학원(INSA de Lyon) 전자공학전공(공학박사)
1998년 4월~1998년 12월 (프)CREATIS 박사후 과정
1999년 3월~현재 영남대학교 전자정보공학부 조교수
관심분야 : 음성, 영상 신호처리, 인공지능, 디지털 워터마킹

정 현 열

1975년 영남대학교 전자공학과(공학사)
1989년 일본 동북대학교 정보공학과(공학박사)
1989년 3월~현재 영남대학교 전자정보공학부 교수
1992년 7월~1993년 7월 미국 CMU Robotics 연구소 객원연구원
1994년 12월~1995년 2월 일본 토요하시기술과학대학 외국인 연구자
2000년 6월~2000년 8월 미국 Qualcomm Inc. 수석 엔지니어
관심분야 : 음성인식, 화자인식, 음성합성 및 DSP 응용분야

교 신 저 자

정 현 열 712-749 경북 경산시 대동 영남대학교 소재관 208호 영남대학교 전자정보공학부