

한국어 분절음 인식을 위한 인식 단위에 대한 연구

A Study on Recognition Units for Korean Speech Recognition

황 영 수*, 송 민 석**, Michael W. Macon***

(Young Soo Hwang*, Min Suck Song**, Michael W. Macon***)

* 관동대학교 전자정보공학과, ** 관동대학교 영어영문과, *** CSLU, OGI, U.S.A.

(접수일자: 2000년 6월 13일; 채택일자: 2000년 7월 26일)

본 연구는 한국어 분절음 인식을 위한 인식단위 설정에 대한 연구이다. 대용량 음성 인식을 수행할 경우, 표준 패턴의 인식 단위를 단어나 음절이 아닌 분절음 단위로 사용하여야 효율적인 음성 인식을 수행할 수 있다. 본 연구는 이와 같은 분절음 인식을 수행하기 위한 연구로서, 인식 단위 설정 변화에 따른 인식 결과를 미국 OGI 연구소의 speech toolkit을 이용하여 검토한다. 인식 단위에 관해서 특히 모음의 경우 철자에 기초한 음소별 인식단위 설정과 현대어 발음에 기초한 인식단위 설정을 비교했으며, 그 결과 발음에 기초해 몇 개의 모음을 통합한 경우가 더 우수한 결과를 보였다. 또한 인식 단위의 설정에 있어서 독립된 분절음으로 설정한 경우보다 앞, 뒤의 소리의 상황을 고려한 바이폰(biphone)을 이용할 경우가 5.7%-25.9%의 향상된 인식 결과를 보였다. 인식 방법에 있어서는 HMM 만을 이용한 방법보다 신경회로망과 HMM을 결합한 인식 방법이 6.1%-7.5%의 더 좋은 인식률을 나타내었다.

핵심용어: 분절음, 바이폰, 인식 단위

투고분야: 음성처리 분야(2.5)

In the case of making large vocabulary speech recognition system, it is better to use the segment than the syllable or the word as the recognition unit. In this paper, we study on the proper recognition units for Korean speech recognition. For experiments, we use the speech toolkit of OGI in U.S.A. The result shows that the recognition rate of the case in which the diphthong is established as a single unit is superior to that of the case in which the diphthong is established as two units, i.e. a glide plus a vowel. And also, the recognition rate of the case in which the biphone is used as the recognition unit is better than that of the case in which the mono-phoneme is used.

Key words: Speech, Biphone, Recognition unit

I. 서 론

디지털 컴퓨터의 응용 기술과 반도체 기술 및 디지털 신호 처리 기술이 급격히 발전함에 따라 음성은 인간과 인간 사이의 의사 소통뿐만 아니라, 인간과 기계 사이의 의사 소통을 위한 매개체로서의 역할이 요구되고 있다. 인간의 가장 자연스러운 정보 교환 매체인 음성을 통하여 기계와 인간이 서로 정확하게 정보를 전달하도록 하는 것을 목표로 하는 음성 인식에 관한 국내의 연구는 어느 정도 성과는 보이고 있으나, 화자에 따른 문제, 음성의 연속성, 음운학적 모호성, 어휘량 문제 등 여러 원인에 의해 자연스러운 음성 인식의 수준에는 못 미치고 있는 실정이다.

음성인식 시스템은 1970년대 초부터 지금까지 활발히 연구되어 왔으며, 대표적인 인식 기법으로는 음성 발생시간 상에서의 패턴 정합에 의해 음성을 인식하는 DP(Dynamic

Programming) 정합 방법[1], 인식 계산량과 메모리량을 적게 하기 위한 데이터 압축 기술을 이용한 벡터 양자화 (Vector Quantization) 기법[2], Markov 모델의 확률적 추정 에 의한 기법을 도입한 HMM(Hidden Markov Model)[3] 과 음성의 인지 과정을 모델화한 인공 신경 회로망[4] 등을 이용한 것들이 있으며, 현재는 위의 기법들을 서로 결합 시켜 인식률 향상을 얻고자 노력하고 있다.

인식률은 상기의 패턴 인식 방법들 외에, 표준 패턴으로 저장하는 음성 인식 단위를 어느 것으로 하느냐에 따라 그 성능이 크게 좌우된다. 상기의 인식 방법들이 전 세계 어느 언어에나 적용될 수 있는 기법들임을 감안한다면, 결국 언어마다 나타나는 인식률의 차이는 한 언어를 위한 인식 단위를 어떻게 설정하느냐에 달려 있다. 따라서 한국어 인식에 중요한 인식 시스템의 요소는 상기의 패턴 방법에 대한 연구보다도 우리말의 인식 단위에 대한 연구가 중요하다.

본 연구에서는 신경 회로망과 HMM을 이용하여, 우리 말 인식 단위의 형태를 변화시켜 우리말 인식 시스템에 적합한 인식 단위를 찾고자 한다. 특히 모음의 경우에 있어서, 철자에 기초한 이론적 음소에 기초한 단위설정과 발음에 기초한 통합적 인식단위 설정의 경우 어느 것이 한국어에 적합한 지를 비교 검토할 것이다.

II. 신경 회로망과 HMM

2.1. 신경회로망

다층 구조의 퍼셉트론은 입력과 출력층 사이에 한 개 이상의 은닉층(hidden layer)이 있는 것으로 단층 퍼셉트론의 여러 가지 단점을 극복할 수 있다. 예를 들어 단층 퍼셉트론은 배타적 OR(exclusive OR) 기능을 수행할 수 없는데 비하여, 2층 구조의 퍼셉트론은 이를 계산할 수 있다. 그러나 다층 퍼셉트론은 이에 대한 효율적 학습 알고리즘이 발견될 때까지 별로 사용되지 않았다.

Rumelhart와 Hinton은 다층 퍼셉트론에 대한 효율적 학습 방법을 개발함으로써 비록 그 방법의 완전성(알고리즘이 항상 최적해에 수렴하는 성질)이 단층 퍼셉트론에서와 같이 명확히 증명되지는 않았지만, 많은 흥미 있는 문제에 성공적으로 응용할 수 있는 계기를 만들었다.

단층 퍼셉트론의 기능은 각 처리 인자(processing element)의 활성화 함수의 비 선형성에 의존한다. 활성화 함수가 선형인 경우, 단층 퍼셉트론으로 다층망(network)의 어떠한 계산 기능도 수행할 수 있다. 그러나 활성화 함수가 비 선형(hard limiting)일 경우, 단층, 2층 및 3층 구조의 퍼셉트론의 계산 기능은 차이를 보인다. 그리고 단층 퍼셉트론은 반 평면(halfplane) 결정 구역을 형성한 반면에, 2층 퍼셉트론은 볼록(convex) 결정 구역을 형성한다. 3층 퍼셉트론은 망사(mesh) 형태의 패턴군들을 분류하는 임의의 결정 구역을 형성할 수 있는 반면, 2층 퍼셉트론은 결코 망사 분류(mesh class) 구역을 분류할 수 없다. 3층 구조의 퍼셉트론이 임의의 결정 구역을 형성하는 데 있어 그 망의 2층 노드의 수는 최대 분리된 구역의 수이며, 1층 노드의 수는 각 2층 노드에 의하여 형성되는 볼록(convex) 구역을 이루는 경계면 총수와 같다.

한편 비선형 활성화 함수가 비 선형이 아니라 S자형(sigmoid)일 경우, 결정 구역이 보통 직선이 아닌 완만한 곡선으로 경계지워짐으로 행태(behavior) 분석이 약간 더 복잡하나 이 경우 오차 역방향 전파(error backward propagation, EBP) 학습 알고리즘을 사용하여 학습시킬 수 있다.

2.2. HMM (Hidden Markov Model)

음성 신호를 시변 랜덤 과정(time-varying random process)이라고 가정하고, 각 음소가 갖는 독특한 특성들을 효과적으로 표현하기 위해서 HMM을 사용한다. HMM은 비슷한 신호 특성들을 대표하는 상태(state)와 상태 간의 천이를 나타내는 상태 천이 행렬(state transition matrix) A, 그리고 각 상태에서의 신호 특성을 나타내는 상태 관찰

행렬(state observation matrix) B로 구성된다. 여기에 초기에 발견될 상태의 확률을 표현하는 초기 상태 벡터(initial state vector) π 를 포함시켜 HMM을 $\lambda = (A, B, \pi)$ 로 표시한다.

상태 천이 행렬 A는 신호의 시변 특성을 나타낸다. HMM이 N 개의 천이 상태로 이루어지면, A 는 NxN 행렬이 되며, 다음과 같이 주어진다.

$$A = \{ a_{ij} \}$$

$$a_{ij} = p(st = i, st + 1 = j) \text{-----} (1)$$

여기에서 st 는 시간 t 에서 발견되는 상태이다. 즉, A 행렬의 각 성분은 행에 해당하는 상태에서 열에 해당하는 상태로 천이할 확률값을 나타낸다. 음성 신호는 파형 특성이 시간에 따라 한쪽 방향으로만 변하기 때문에, HMM의 상태 천이도 마찬가지로 한 방향으로만 흐르도록 규정할 수 있다. 이것을 좌우 모델(left-right model)이라 하며, 현 상태보다 앞에 있는 상태로 천이할 수 없도록 한다. 그러면 $a_{ij} = 0, (i > j)$ 가 되어, A 행렬은 상위 삼각행렬(upper triangular matrix)이 된다.

상태 천이시 너무 많이 뛰어 넘는 것을 방지하기 위해 최대 천이지 p를 규정하면, 행렬 A는 넓이 p를 갖는 대행렬(band matrix)이 된다.

$$\begin{bmatrix} a_{00} & a_{01} & 0 & 0 & \dots & 0 \\ 0 & a_{11} & a_{12} & 0 & \dots & 0 \\ 0 & 0 & 0 & a_{22} & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 1 \end{bmatrix}, (p=2) \text{-----} (2)$$

각 상태에서 관찰되는 파형들이 가정상적(quasi-stationary) 특성을 나타낸 것이 밀도 행렬 B이다. 이때 관찰값을 양자화하느냐, 연속적인 형태로 하느냐에 따라서 HMM의 종류를 분류할 수 있다. 전자의 경우를 이산 밀도 HMM(Discrete Density HMM)이라 하고, 후자를 연속 밀도 HMM(Continuous Density HMM)이라고 한다.

이산 밀도 HMM에서는 D 차 관찰값을 L 개의 값을 갖는 코드북(codebook)으로 벡터 양자화하고, 이를 통해 얻어지는 코드워드를 HMM의 관찰값으로 생각한다. 이와 같은 경우 확률 밀도 행렬 B의 각 항은 N 번째 상태에서 L 번째 코드워드를 관찰할 확률을 나타낸다. 즉,

$$B = \{ b_{j(l)} \}$$

$$b_{j(l)} = p(st = j, ot = l) \text{-----} (3)$$

이며, NxL 행렬로 표현된다.

위의 경우와 달리, 관찰값을 D 차원의 연속적인 값 그대로 사용하는 경우가 연속 밀도 HMM이다. 이때 확률 밀도 행렬 B는 각 상태의 확률 밀도를 연속적인 혼합 모수 밀도(mixture parametric density)로 표시하며, 일반적으로 F 개의 가우시안 분포의 합으로 나타낸다.

한편 이산 HMM과 연속 HMM을 결합한 것을 반연속 HMM(Semi-Continuous HMM)이라 하며, 이 경우는 벡터

양자화 코드워드를 가우시안 분포의 평균치들로 생각하며, 각 분포의 공분산 행렬의 대각선 값들을 코드북에 포함 시키게 된다. 즉, 확률 밀도 행렬 B는 상태 j에서 l 번째 코드워드에 해당하는 가우시안 성분을 발견할 상대적인 크기가 되므로, b_{jl}은 연속 밀도 HMM의 c_{jl}와 같은 역할을 한다. 그러면 상태 j에서 관찰값 α_t를 발견할 확률은

$$p_j(\alpha_t) = \sum_l b_{jl} p(\alpha_t | \mu_{jl}, \Sigma_{jl}) \dots \dots \dots (4)$$

로 주어진다.

이와같은 여러 HMM을 이용하여 음성의 학습 데이터를 잘 표현하기 위해서는, HMM의 모수 재추정(parameter reestimation) 과정이 필요하다. 이것은 모수가 주어졌을 때, 관찰열을 발견할 확률을 반복적으로 최대화시키는 것으로서 EM(Expectation Maximization) 알고리즘이라 한다. 또한 주어진 HMM 모수들로부터 하나의 관찰열에 대응되는 가장 적합한 상태열을 찾는 방법으로 비터비(Viterbi) 알고리즘이 있다.

III. 한국어 인식단위와 인식 시스템

3.1. 한국어 인식단위

먼저 한국어에서 사용되는 자음 체계를 보면 모두 19개의 음소가 사용되고 있다.

자음: ㄱ ㄴ ㄷ ㄹ ㅁ ㅂ ㅅ ㅇ ㅈ ㅊ ㅋ ㅌ ㅍ ㅎ ㅊ
ㅌ ㅍ ㅈ ㅊ

이 자음들에 대해서는 음소 하나 당 각기 하나의 인식 단위가 설정될 수 있다. 이 인식단위들을 자음 체계의 입장에서 구성한 표가 표 1이며, 표 2에 인식 단위에 대한 Worldbet 기호를 나타내었다.

표 1. 자음 체계
Table 1. Korean Consonant.

	평음	격음	경음	공명음
연구개음	ㄱ	ㅋ	ㄲ	ㅇ
치경음	ㄷ	ㅌ	ㄸ	ㄴ
양순음	ㅂ	ㅍ	ㅃ	ㅁ
치경구개음	ㅈ	ㅊ	ㅉ	ㄹ
치경마찰음		ㅅ	ㅆ	
후두음		ㅎ		

표 2. 자음체계의 Worldbet 표현
Table 2. Worldbet corresponding to Korean Consonant.

음소	Worldbet	음소	Worldbet	음소	Worldbet	음소	Worldbet
ㅂ	p	ㅃ	t*	ㅅ	ch	ㅁ	m
ㅍ	ph	ㄱ	k	ㅋ	c*	ㄴ	n
ㅍ*	p*	ㅋ	kh	ㅅ	s	ㅇ	N
ㄷ	t	ㄲ	k*	ㅆ	s*	ㄹ	l
ㅌ	th	ㅈ	c	ㅊ	h		

그리고 자음 중 파열음과 파찰음, 마찰음(ㄱ, ㅋ, ㄲ, ㄷ, ㅌ, ㅈ, ㅊ, ㅉ, ㅆ, ㅅ, ㅆ, ㅈ, ㅊ, ㅉ, ㅆ)이 음절의 종성 위치에 올 경우, 조음 위치에 따라 각기 'ㄱ, ㄷ, ㅂ'로 중화(neutralized)되고, 초성과는 다른 음향적 특성을 보이므로 별도의 인식단위로 설정하였다. 또한 'ㄷ'과 'ㅎ'의 경우는 나타나는 환경에 따라 뚜렷한 음향적 특성의 차이를 보이므로 기본 음소 외에 환경에 따라 추가로 별도의 인식단위를 설정하였다. 따라서 한국어 인식을 위해 설정된 자음 인식단위는 총 24개이다.

한국어의 모음으로 사용되는 음소는 아래의 21개이다.

단모음: ㅏ ㅑ ㅓ ㅕ ㅗ ㅛ ㅜ ㅠ ㅡ ㅟ ㅞ ㅠ
이중모음: ㅘ ㅙ ㅚ ㅜㅝ ㅝㅞ ㅞㅟ ㅠㅡ ㅢㅣ

그러나 실제 모국어 화자들의 발음을 살펴보면, 단모음으로 분류된 'ㅛ'의 경우 현대어에서는 거의 이중모음으로 변화되었고, 단모음 'ㅞ, ㅠ'의 경우는 50이하의 젊은 층에서는 거의 'ㅞ'로 통합이 되었다. 이중모음 'ㅝ'의 경우도 거의 'ㅞ'로 통합이 되었고, 'ㅠ, ㅡ'의 경우도 'ㅛ'와 함께 하나로 통합이 되었다.

본 연구에서는 이론적 음소에 기초한 21개의 음소를 인식단위로 설정한 경우와 실제 발음에서 통합된 모음은 하나로 설정하여 17개의 인식단위를 설정한 경우를 HMM 방식에 의해 인식률을 비교해 보았고 4-3절 표 4에 결과가 제시되어 있듯이 발음에 근거해 통합된 모음을 하나로 설정한 경우가 더 우수한 것으로 나타났다. 따라서 인식 방법을 실험한 2차 실험부터는 모음을 위한 인식단위로 17개의 인식단위를 사용하였다. 이 17개 모음에 대한 Worldbet 표기가 표 3에 제시되어 있다.

표 3. 모음의 Worldbet 표기
Table 3. Worldbet corresponding to Korean Vowel.

철자	음소 (IPA)	Worldbet	철자	음소 (IPA)	Worldbet	철자	음소 (IPA)	Worldbet
ㅏ	a	a	ㅑ	e	e	ㅓ	wa	ua
ㅕ	ɔ	&	ㅗ	ja	ia	ㅛ	wɔ	u&
ㅓ	o	o	ㅕ	jɔ	i&	ㅜ, ㅠ, ㅠ	we	ue
ㅕ	u	u	ㅟ	jo	io	ㅛ	wi	ui
ㅡ	i	ix	ㅠ	ju	iu	ㅛ	ii	ixi
ㅠ	i	i	ㅡ	je	ie			

3.2. 음성 인식 시스템

본 연구에서 사용한 음성인식 시스템의 구성도를 나타낸 것이 그림 1이다. 그림 1의 음성인식 시스템은 신경회로망과 HMM을 결합한 방법이다. 인식 시스템의 세 번째 단계에서 입력 음성의 프레임을 이용하여 분절음 단위 인식을 수행하는 단계로서, 이 단계에서는 신경회로망을 이용한다. 이와 같이 세 번째 단계에서 분절음 단위 인식을 수행한 후, 네 번째 단계에서는 Viterbi 방법을 이용하여 단어 인식을 수행하게 된다.

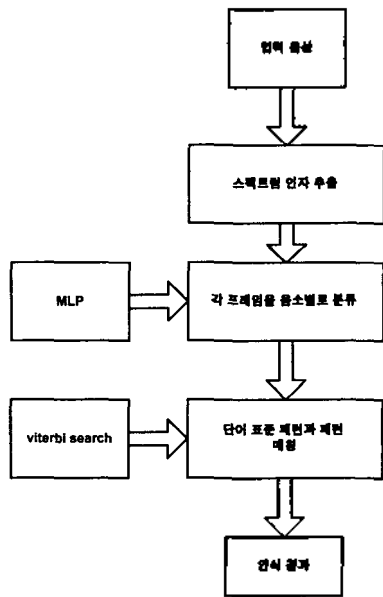


그림 1. 음성인식 시스템
Fig. 1. Block diagram of speech recognition used in this paper. (OGI speech tool kit).

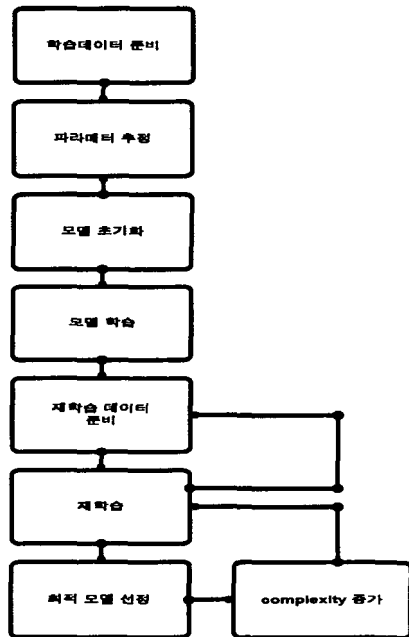


그림 2. HMM 학습 과정 블록도
Fig. 2. Block diagram of Training Process of HMM.

본 연구에서 사용한 OGI의 패턴 유사도 결정 방법은 신경회로망, HMM과 이 두 방법을 결합한 방법을 비교하였다. 이 중 HMM을 이용한 학습 블록도를 그림 2에 나타내었다.

그림 2의 1 단계 데이터 준비 과정에서는, 표준 패턴용 데이터에 III-1절에서 설명한 인식단위를 수작업으로 레이블링하는 과정이다. 3 단계 모델 초기화 과정에서는, 2

단계에서 구한 표준 패턴의 특징 파라미터들을 벡터 양자화하여 각 인식단위 모델을 초기화한다. 4 단계에서는 EM(Expectation/maximization)알고리즘을 이용하여 HMM 모델을 학습한다. 5 단계에서는 6 단계에서의 재학습 과정에 필요한 학습 데이터의 단어 목록을 구성한다. 6 단계에서는 5 단계에서 구한 단어 목록에 수작업이 아닌 자동으로 인식단위 구간 추출을 수행한 후, HMM 모델을 재추정하고, 7 단계에서는 재추정된 여러 HMM 모델 중 최적의 인식 결과를 갖는 HMM 모델을 선정한다.

신경회로망을 이용한 학습 시에는 학습 데이터를 분류하기 위하여 벡터양자화 방법을 사용한다. 모노폰(monophone)이 아닌 바이폰(biphone)을 이용한 음성 인식을 수행할 경우에는 그림 3에 나타낸 것과 같이, 모음에는 3 영역, 자음에는 2 영역으로 구분하여 조합 형태의 모델을 이용하였다.

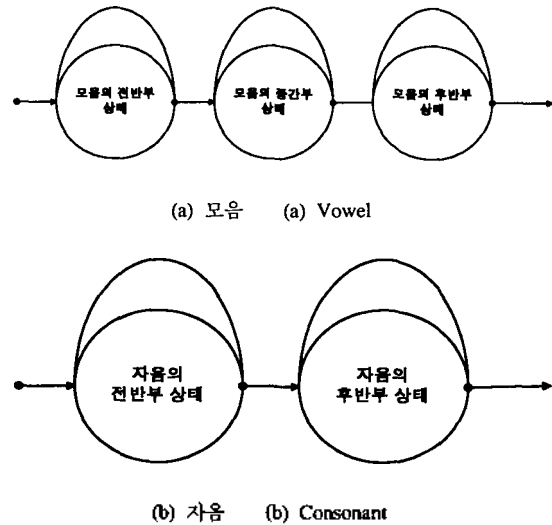


그림 3. biphone구성을 위한 모음, 자음부의 분류
Fig. 3. Classification of Vowel and Consonant for making biphone.

IV. 실험 및 결과 고찰

4.1. 실험 데이터

실험에 사용된 데이터는 격리 단어 452개를 9명이 2번씩 발성한 데이터를 이용하였다. 9명중 4명(남자 2명, 여자 2명)이 발성한 데이터를 학습에, 나머지 5명(남자 4명, 여자 1명)의 데이터와 학습에 포함된 화자가 다른 시기에 발성한 데이터를 인식실험에 사용하였다. 또한 상기 9인 외의 남성 1인 여성 1인이 발성한 다른 데이터(학습에 사용한 단어 외의 데이터)를 인식실험에 사용하였다. 이 데이터들은 16KHz, 16bit로 샘플링(sampling)하였으며, 인식 파라미터는 13차 멜 켈스트럼(Mel cepstrum) 계수를 기본으로 평균값을 뺀 것, 1, 2차 시간 미분 값을 더한 39개의 파라미터를 학습과 인식실험에 사용하였다.

4.2. 인식 시스템

본 논문에서 사용한 HMM은 일반적인 HMM으로서, 상태수 5개(3개의 관측 상태, 1개의 entry와 1개의 exit)로서 좌에서 우방향(left-to-right) 모델을 각 인식단위별로 구성하였다. 또한 하이브리드(hybrid) 시스템에서 사용된 HMM은 상태수를 3개(1개의 관측상태, 1개의 entry와 1개의 exit)를 사용하였으며, 신경회로망은 1개의 은닉층을 갖는 MLP구조를 사용하였다.

인식 단위를 바이폰(biphone)으로 사용할 경우, 모음(나, 키, 누, 터, 이, 케)과 이중모음(크, 키, 쓰, 프, 케, 과, 거, 니, 구, ...)에서는 3 영역(전반부, 정상 상태, 후반부)으로 자음(ㄱ, ㄴ, ㄷ, ㄹ, ㅁ, ㅂ, ㅅ, ㅇ, ㅈ, ㅊ, ㅋ, ㅌ, ㅍ, ㅎ, ㄲ, ㄴ, ㄷ, ㄹ, ㅁ, ㅂ, ㅅ, ㅇ, ㅈ, ㅊ, ㅋ, ㅌ, ㅍ, ㅎ)에서는 2 영역(전반부, 후반부)으로 구분하여, 각 모델들의 조합 갯수 만큼의 HMM 모델을 설정하였다.

4.3. 실험 결과

표 4와 그림 4에 HMM을 이용한 인식 실험 결과를 나타내었다.

표 4. HMM을 이용한 인식 결과

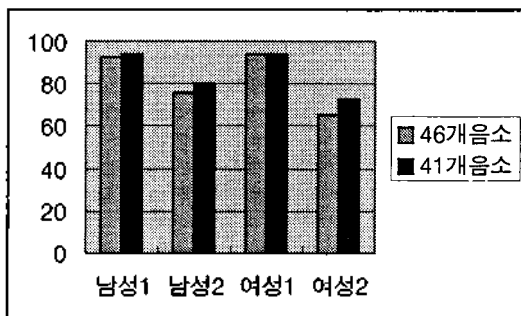
Table 4. Recognition Result using HMM.

(a) 학습에 포함된 화자 데이터 결과
(a) Result using Training Speakers

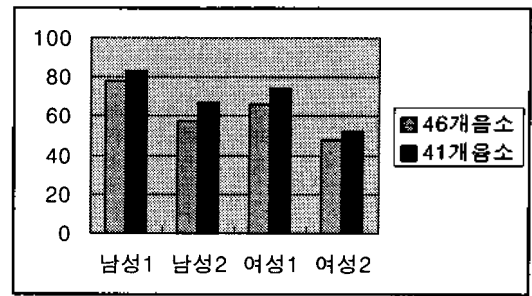
	남성 1	남성 2	여성 1	여성 2
46개 음소	92.7	75.4	93.8	65.3
41개 음소	94.0	80.1	94.0	72.6

(b) 학습에 포함되지 않은 화자 데이터 결과
(b) Result using Non-Training Speakers

	남성 3	남성 4	남성 5	남성 6	여성 3
46개 음소	78.1	57.1	52.7	65.9	47.8
41개 음소	83.4	66.6	60.4	74.1	52



(a) 학습에 포함된 화자 데이터 결과



(b) 학습에 포함되지 않은 화자 데이터 결과

그림 4. HMM을 이용한 인식 결과

Fig. 4. Recognition Result using HMM.

표 4에 나타난 것 같이 41개 음소를 이용한 결과가 46개를 이용한 인식 결과보다 학습시 포함된 화자나 학습시 포함되지 않은 데이터 모두 더 우수한 결과를 보이고 있다. 이 결과에 따라 Hybrid 시스템에서는 41개 음소를 이용해 인식실험을 수행했으며, 그 결과를 인식 단위별로 표 5와 그림 5에 나타내었다.

표 5. 인식단위 설정에 따른 인식 결과

Table 5. Recognition Result according to recognition unit.

phone	남성1	남성2	남성3	남성4	남성5	남성6	여성1	여성2	여성3
mono	89.6	73.9	73.9	60.8	50.4	70.6	77.7	75.9	66.6
bi	97.6	97.3	91.6	79.4	76.3	81.6	88.3	89.8	72.3

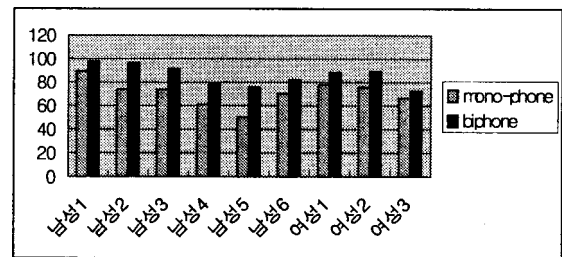


그림 5. 인식단위 설정에 따른 인식 결과

Fig. 5. Recognition Result according to recognition unit.

표 5에 나타난 것 같이 특정 화자의 데이터에 관계 없이 인식단위를 모노폰(mono-phone)으로 설정한 경우보다 바이폰(biphone)으로 설정할 경우, 5.7%-25.9%의 인식을 상승 효과를 보여주고 있다.

표 6에는 학습에서 사용하지 않은 단어들로 구성된 데이터를 이용한 인식 결과를 나타내었다. 표 6에서 HMM의 인식 결과는 재학습을 통하여 제일 최적화된 결과를 나타낸 인식 결과이다. 표에 나타난 것과 같이 바이폰을 사용한 하이브리드(hybrid) 시스템의 결과가 모노폰을 사용한 신경회로망 시스템의 경우보다 42.2%까지 더 높은 인식을 보이고 있다.

표 6. 학습에 포함되지 않은 단어의 인식률
Table 6. Recognition Result using words not included in training words.

	biphone (hybrid)	mono-phone (신경회로망)	mono-phone (HMM)
남	77.5	35.3	70
여	88.8	51.0	82.7

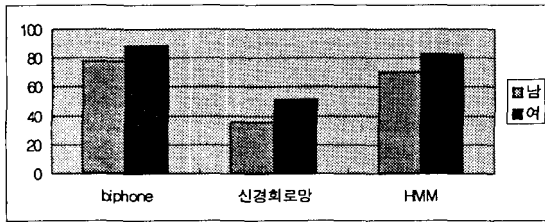


그림 6. 학습에 포함되지 않은 단어의 인식률
Fig. 6. Recognition Result using words not included in training words.

표 4, 표 5, 표 6에 나타낸 것과 같이 음성인식에서 사용되는 인식단위는 이론적 음소에 근거해 설정한 경우보다 발음에 근거해 설정한 경우에 더 높은 인식률을 얻을 수 있었고, 인식 방법에 관계없이 모노폰(mono-phone)으로 설정한 방법보다는 바이폰(biphone)으로 설정한 인식 결과가 상당히 우수한 결과를 보여주었다.

V. 결 론

본 논문은 한국어 음성인식 시스템을 구성할 경우, 인식 시스템의 패턴 매칭부의 인식 방법의 변동에 따른 인식 시스템의 인식률 향상이 아닌 입력되는 음성 자체 즉, 한국어의 특성을 알고 그에 따른 인식단위 설정에 의한 음성인식 시스템의 성능 향상을 얻고자, 인식단위에 따른 한국어 음성인식 결과를 실험 고찰한 것이다.

본 논문에서 사용한 인식단위는 모음 21개의 음소 중 '애, 에', '내, 니, 네', '해, 헤'를 같은 인식단위로 설정하여 17개의 모음 인식단위 모델을 설정하였으며, 자음에서는 19개의 음소에 초성과 중성의 위치에 따라 그 음향적 특성이 다르게 나타나는 'ㄱ, ㅋ, ㆁ', 'ㄷ, ㅌ, ㄴ, ㄷ, ㅌ, ㄴ', 'ㄹ, ㄹ'을 위한 3개의 인식단위를 추가하여 총 22개의 인식단위를 설정하였다. 따라서 최종 수행된 인식단위의 총 수는 41개이다.


실험 결과, 모음의 경우 위와 같이 17개로 통합하여 인식단위를 설정한 결과가 이론적 음소에 근거해 인식단위를 설정한 결과보다 더 우수한 인식률을 나타내었다. 또한 어느 인식 방법에 관계없이 모노폰(mono-phone)으로 인식단위를 설정하는 것보다 바이폰(biphone)을 이용한 결과가 우수한 것을 알 수 있었고, 제일 좋은 인식 결과는 HMM과 신경회로망을 결합하여 바이폰을 인식단위로 이용한 인식기에서 얻을 수 있었다.


향후 한국어 음성인식에 적합한 인식단위에 대한 연구는 음향적 특성에 따라 인식단위를 변화시켜가며 계속 실험해 한국어에 적합한 최적의 인식단위 세트를 설정해야 하며, 한국어 음성 인식기뿐만 아니라 합성기에 최적인 음성 구조에 대한 연구도 병행해 나아갈 것이다.

참 고 문 헌

1. H. Sakoe, "Two-Level DP matching-dynamic programming based pattern matching algorithm for connected word recognition," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-27, pp. 588-595, Dec. 1979.
2. Y. Linde, A. Buzo and R. M. Gray, "An algorithm for vector quantizer design," IEEE Trans. on Com, Vol. COM-28, Jan., pp. 84-95, 1980.
3. L. R. Rabiner and B. H. Juang, "An Introduction to Hidden Markov Models," IEEE ASSP Mag., Jan. 1986.
4. Y. H. Pao, Adaptive Pattern Recognition and Neural Networks, Addison-Wesley Pub. Co., 1989.
5. A. J. Viterbi, "Error Bounds for Conventional Codes and an Asymptotically Optimal Decoding Algorithm," IEEE Trans. Inf. Theory, Vol. IT-13, pp. 260-269, 1967.
6. J. Schalkwyk, P. Hosom, Ed Kaiser and K. Shobaki, "CSLU-HMM: The CSLU Hidden Markov Modeling Environment," CSLU in OGI, Feb, 1999.
7. J. P. Hosom, R. Cole, M. Party, J. Schalkwyk, Y. Yan and W. Wei, "Training Neural Network for Speech Recognition," CSLU in OGI, Feb, 1999.
8. 황영수, 송민석, "한국어 음성 인식을 위한 Biphone 구성을 위한 기초 연구," 한국음향학회 하계학술대회, July, 2000.

▲ 황 영 수
한국음향학회지 18권 1호 참조

▲ 송 민 석

 관동대학교 영어영문과 교수
 한국의국어 대학교 영어과 졸업, 1983
 한국의대 대학원 영어과 문학석사, 1987
 한국의대 대학원 영어과 문학박사, 1992

▲ Michael W. Macon

 Assistant Professor, Dept. of ECE / Dept. of CSE
 Oregon Graduate Institute of Science & Technology
 Ph.D., Electrical Engineering, Georgia Institute of Technology, 1996.
 M.S.E.E., Georgia Institute of Technology, 1993.
 B.E.E., University of Dayton, 1991.