2000년도 한국음향학회 학술발표대회 논문집 제19권 제1(s)호

한국어 음성 인식용 biphone 구성을 위한 기초 연구

황 영 수*, 송 민 석**

*관동대학교 전자정보공학과.**관동대학교 영어영문과

The Basic Study on making biphone for Korean Speech Recognition

YoungSoo Hwang*, Song, Minsuck**

Division of Information Electronics Engineering,Kwan Dong University

Dept. of English Language and Literature,Kwan Dong University

E-mail : hysoo@mail.kwandong.ac.kr, mmsong@mail.kwandong.ac.kr

ABSTRACT

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 basis of making biphone 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.

I. 서 론

디지탈 컴퓨터의 응용 기술과 반도체 기술 및 디지탈 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 발성 시간 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 추정에 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 결합시켜 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, 우리말 인식 단위의 형태를 변화시켜 우리말 인식 시스템에 적합한 바이폰 형태를 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. 한국어 인식단위와 인식 시스템

II-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
ㄲ ㄸ ㅃ ㅉ ㅆ

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

표 1. 자음 체계

	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

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

그리고 자음 중 파열음과 파찰음, 마찰음(ㄱ, ㅋ, ㄲ, ㄷ, ㅌ, ㅈ, ㅊ, ㅅ, ㅆ, ㅎ, ㅂ, ㅍ)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'와 '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 ㅓ ㅗ ㅜ ㅡ ㅣ ㅔ ㅐ ㅚ
이중모음: ㅑ ㅕ ㅛ ㅠ ㅒ ㅖ ㅘ ㅝ ㅟ ㅙ ㅞ ㅢ

그러나 실제 모국어 화자들의 발음을 살펴보면, 단모음으로 분류된 'ㅚ'의 경우 현대어에서는 거의 이중모음으로 변화되었고, 단모음 'ㅔ, ㅐ'의 경우는 50이하의 젊은 층에서는 거의 'ㅔ'로 통합이 되었다. 이중모음 'ㅒ'의 경우도 거의 'ㅖ'로 통합이 되었고, 'ㅙ, ㅞ'의 경우도 'ㅚ'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

철자	음소(IPA)	Worldbet	철자	음소(IPA)	Worldbet
ㅏ	a	a	ㅔ, ㅐ	e	e
ㅓ	ə	&	ㅑ	ja	ia
ㅗ	o	o	ㅕ	jə	i&
ㅜ	u	u	ㅛ	jo	io
ㅡ	ɨ	ix	ㅠ	ju	iu
ㅣ	i	i	ㅖ, ㅒ	je	ie

철자	음소(IPA)	Worldbet
ㅘ	wa	ua
ㅝ	wə	u&
ㅚ, ㅙ, ㅞ	we	ue
ㅟ	wi	ui
ㅢ	ɰi	ixi

II-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.

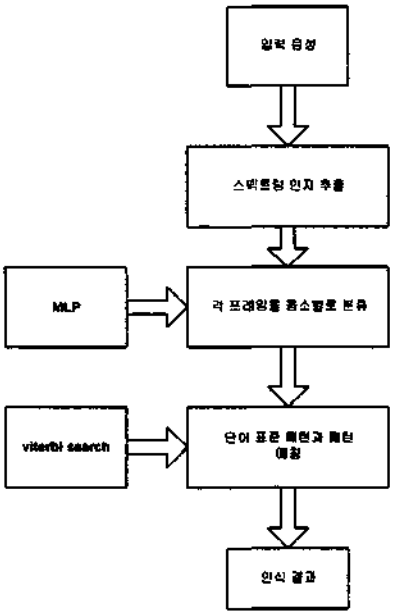

그림 1. 음성인식 시스템

III. 실험 및 결과 고찰

III-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 파라메터는 13차 멜 켑스트럼(Mel cepstrum) 계수를 기본으로 평균값을 뺀 것과, 1, 2차 시간 미분 값을 더한 39개의 파라메터를 학습과 인식실험에 사용하였다.

III-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(ㅏ ㅓ ㅗ ㅜ ㅡ ㅣ ㅐ)과 이중모음(ㅑ ㅕ ㅛ ㅠ ㅖ ㅘ ㅝ ㅟ ㅢ)에서는 3 영역(전반부, 정상 상태, 후반부)으로 자음(ㄱ ㄴ ㄷ ㄹ ㅁ ㅂ ㅅ ㅇ ㅈ ㅊ ㅋ ㅌ ㅍ ㅎ ㄲ ㄸ ㅃ ㅉ ㅆ)에서는 2 영역(전반부, 후반부)으로 구분하여, 각 모델들의 조합 갯수 만큼의 HMM 모델을 설정하였다.

III-3. 실험 결과

표 4 에 HMM을 이용한 인식실험 결과를 나타내었다.

표 4. HMM을 이용한 인식 결과

(a) 학습에 포함된 화자 데이터 결과

	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

	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

표 4에 나타낸 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

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

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 인식률 상승 효과를 보여주고 있다.

IV. 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 'ㅐ, ㅔ', 'ㅖ, ㅚ, ㅙ', 'ㅒ, ㅞ'를 같은 인식단위로 설정하여 17개의 모음 인식단위 모델을 설정하였으며, 자음에서는 19개의 음소에 초성과 종성의 위치에 따라 그 음향적 특성이 다르게 나타나는 'ㄱ, ㅋ, ㄲ', 'ㄷ, ㅌ, ㅈ, ㅊ, ㅉ, ㅅ, ㅆ, ㅎ', 'ㅂ, ㅍ'를 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 결과가 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.

참고문헌

[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.