한국어 인식을 위한 인식 단위와 학습 데이터 분류 방법에 대한 연구
(A Study on Recognition Units and Methods to Align Training Data for Korean Speech Recognition)

황 영 수
Youngsoo Hwang

요 약

본 연구는 한국어 분절음 인식을 위한 인식 단위 설정과 학습시 학습 데이터 분할 방법에 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 이와같은 분절음 인식을 수행하기 위한 연구로서, 인식 단위 설정 변화와 학습시 학습 데이터 분할 방법에 따른 인식 결과를 미국 OGI 연구소의 speech toolkit을 이용하여 검토한다.

인식 단위에 관해서 특히 모음의 경우 철자에 기초한 음소별 인식 단위 설정과 현대어 발음에 기초한 인식 단위 설정을 비교했으며, 그 결과 발음에 기초해 몇 개의 모음을 통합한 경우가 더 우수한 결과를 보였으며, 학습 데이터 분할 방법에 따른 인식 결과는 손으로 분할한 방법이 자동 분할 방법보다 약 2-3%의 인식 향상을 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.

Abstract

This is the study on recognition units and segmentation of phonemes. 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 and segmentation of phonemes 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 recognizer using manually-aligned training data is a little superior to that using automatically-aligned training data. 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 : recognition unit, phoneme, biphone, segmentation

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 발성시간 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 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. 또한 학습시의 음소 분류 방법 중 손으로 분류하는 방법과 자동 분류하는 방법의 인식률 변화를 살펴보는 데에 그 목적이 있다.

II. 한국어 인식 단위

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, 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 1에 제시되어 있다.

표 1. 모음의 Worldbet 표기

Table 1. Worldbet corresponding to Korean Vowel

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	ii	ixi

또한 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 2이며, 표 3에 인식 단위에 대한 Worldbet 기호를 나타내었다.

표 2. 자음 체계

Table 2. 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 3. 자음체계의 Worldbet 표현

Table 3. Worldbet corresponding to Korean Consonant

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 인식 단위를 설정하였다. 따라서 한국어 인식을 위해 설정된 자음 인식단위는 총 24개이다. 또한 'ㄱ', 'ㄷ', 'ㅂ', 'ㄹ', 을 위치에 따라 한 개씩의 음소를 첨가하여 인식 단위를 28개로 할 경우와, 'ㅅ', 'ㅁ', 'ㄴ', 'ㄹ', 'ㅇ'을 각각 위치에 따라 2개, 3개, 3개, 3개, 2개의 음소로 구분하여 인식 단위를 35개로 하여 각각 실험을 수행하였다

III. 음성 인식 시스템

1. 본 논문에서 사용한 음성 인식 시스템

본 연구에서 사용한 음성 인식 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.

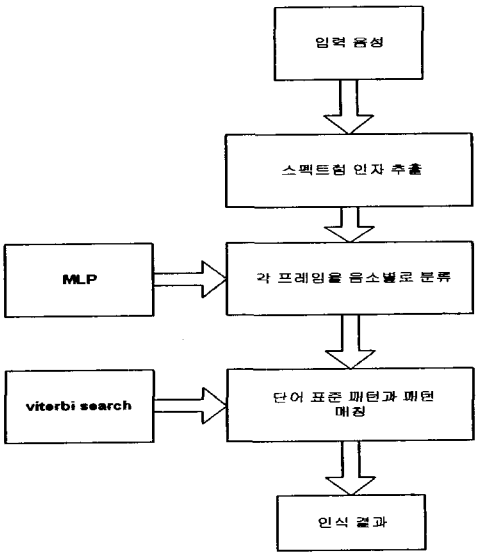
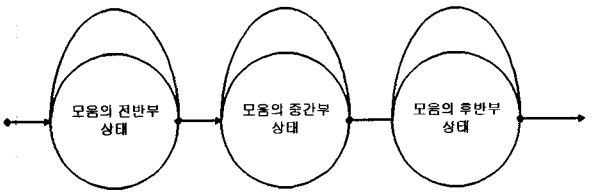


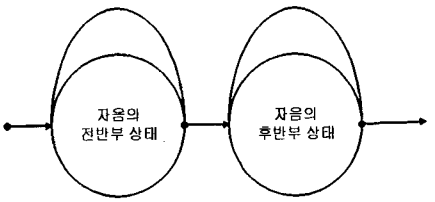
그림 1. 본 논문에서 사용한 음성 인식 시스템
Fig 1. Block Diagram of Speech Recognition System in This Paper (OGI Speech Tool Kit)

본 연구에서 사용한 OGI의 패턴 유사도 결정 방법은 신경회로망, HMM과 이 두 방법을 결합한 방법을 비교하였다. 이 방법의 1 단계 데이터 준비 과정은, 표준 패턴용 데이터에 본 논문에서 설명한 인식단위를 수작업으로 레이블링하는 과정이다. 3 단계 모델 초기화 과정에서는, 2 단계에서 구한 표준 패턴의 특징 파라메터들을 벡터 양자화하여 각 인식단위 모델을 초기화한다. 4 단계에서는 EM(Expectation/maximization)알고리듬을 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(mono-phone)이 아닌 바이폰(biphone)을 이용한 음성 인식을 수행할 경우에는 그림 2에 나타낸 것과 같이, 모음에는 3 영역, 자음에는 2 영역으로 구분하여 조합 형태의 모델을 이용하였다. 모음을 2 영역으로 구분 한 이유는 모음 앞 부분의 연결부와 모음 후반부의 연결부 그리고 모음의 정상상태를 고려한 것이고, 자음은 자음의 시간길이 때문에, 정상상태를 고려하지 않고 앞, 뒤의 연결부만 고려하였다.



(a) 모음
(a) Vowel



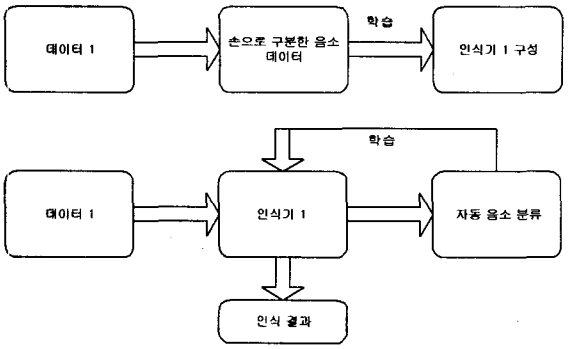
(b) 자음
(b) Consonant

그림 2. biphone을 이용한 모음, 자음부의 상태도
Fig 2. Classification of Vowel and Consonant for making biphone

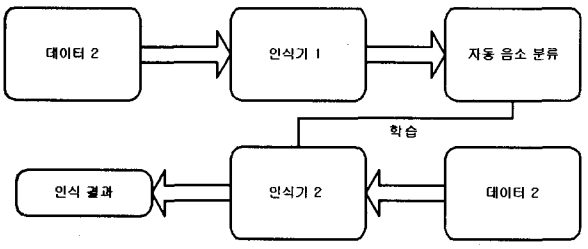
2. 음소 분류 방법

일반적으로 학습시 학습 데이터를 손으로 분류하여 수행하는 것이 자동 분류하는 것보다 성능이 우수하다고 알려져 있지만, 실제 데이터의 결과에 따른 수치 비교를 하기 위하여, 본 연구에서는 상기의 음소 분류와 이에 대한 학습시 음소 분류 방법에 따른 인식 결과를 알아보고자 한다. 손으로 음소를 분류하는 방법과 자동으로 음소

분류하는 방법에 따른 인식률 결과를 살펴보기 위하여, 그림3에 나타낸 것과 같은 블록도를 이용하였다.



(a) 데이터 1을 이용한 분류
(a) Alignment by using data 1



(b) 데이터 2를 이용한 분류
(b) Alignment by using data 2

그림 3. 음소 분류에 따른 학습 방법
Fig 3. Training Methods According to the Alignment of Phoneme

일반적으로 학습시 학습 데이터를 손으로 분류하여 수행하는 것이 자동 분류하는 것보다 성능이 우수하다고 알려져 있지만, 실제 데이터의 결과에 따른 수치 비교를 하기 위하여, 본 연구에서는 상기의 음소 분류와 이에 대한 학습시 음소 분류 방법에 따른 인식 결과를 알아보고자 한다. 손으로 음소를 분류하는 방법과 자동으로 음소 분류하는 방법에 따른 인식률 결과를 살펴보기 위하여, 그림3에 나타낸 것과 같은 블록도를 이용하였다.

그림 3에서 자동 분류 방법은 CSLU Toolkit의 forced 분류 방법[6]을 사용하였다.

그림 3(a)는 손으로 음소를 구분한 후, 이 데이터를 이용하여 인식기 1을 구성하여 인식 실험을 수행하고, 두 번째로는 이 인식기를 이용하여, 동일 학습 데이터를 자동으로 분류한 후, 이 데이터를 이용하여 인식기를 구성한 후 인식 실험을 수행하였다. 그림 3(b)는 그림 3(a)에서 구성한 인식기를 이용하여 다른 학습 데이터를 자동으로 분류한 후, 이 자동 분류된 데이터를 이용하여 그림 3(a) 와 다른 인식기를 구성한 후 자동 분류된 데이터와 손으로 분류된 데이터를 사용하여 인식 실험을 수행하였다.

Ⅳ. 실험 및 결과 고찰

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 캡스트럼(Mel cepstrum) 계수를 기본으로 평균값을 뺀 것과, 1, 2차 시간 미분 값을 더한 39개의 파라메터를 학습과 인식실험에 사용하였다.

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 ㅔ)과 이중모음(ㅑ ㅕ ㅛ ㅠ ㅖ ㅘ ㅝ ㅚ 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 갯수만큼의 HMM 모델을 설정하였다.

3. 실험 결과

표 4에 인식기로 HMM을 이용하고, 이중 모음의 변화에 따른 인식 실험 결과를 나타내었다.

표 4. 이중 모음에 따른 인식 결과(%)
Table 4. Recognition Result(%) by using Diphthong

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

	남성 1	남성 2	여성 1	여성 2
구분(O)	92.7	75.4	93.8	65.3
구분(X)	94.0	80.1	94.0	72.6

(b) 학습에 포함되지 않은 화자 데이터 결과

(b) Result using Non-Training Speakers' Data

	남성 3	남성 4	남성 5	남성 6	여성 3
구분(O)	78.1	57.1	52.7	65.9	47.8
구분(X)	83.4	66.6	60.4	74.1	52

표 4 에 나타낸 것 같이 이중 모음을 구분하지 않은 인식 결과가 이중 모음을 구분하여 인식한 결과보다 학습시 포함된 화자나 학습시 포함되지 않은 데이터 모두 더 우수한 결과를 보이고 있다. 이 결과에 따라 Hybrid 시스템에서는 이중 모음을 구분하지 않은 음소를 이용해 인식 실험을 수행했으며, 그 결과를 인식 단위별로 표 5 에 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

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

표 6. 학습에 포함되지 않은 단어의 인식률(%)
Table 6. Recognition Result(%) using words not included in Training Data

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

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.

또한 자음 개수에 따른 인식률의 결과를 표 7에 나타내었다.

표 7에 나타낸 자음에 따른 인식률 변화는 학습 화자 포함된 결과나 학습 화자를 포함 시키지 않은 결과 모두 큰 변화를 보여주지 못하고 있으며, 이는 실험 데이터의 변화에 따른 결과라고 사료된다.

표 7. 자음에 따른 인식률(%)

Table 7. Recognition result(%) according to the number of consonants

(a)학습에 포함된 화자 데이터 결과

(a) Recognition Result using Training Speakers' Data

음소 수	남성 1	남성 2	여성 1	여성 2
41	94.0	80.1	94.0	72.6
45	95.8	82.1	93.4	70.8
52	96.2	80.1	93.0	71.7

(b) 학습에 포함되지 않은 화자 데이터 결과

(b) Recognition Result using Non-Training Speakers' Data

음소 수	남성 3	남성 4	남성 5	남성 6	여성 3
41	83.4	66.6	60.4	74.1	52.0
45	83.4	66.4	60.2	76.1	60.2
52	84.7	68.1	60.8	76.6	54.0

학습 방법에 따른 인식 실험에 사용된 데이터는 상기 실험에 사용한 데이터와 동일하며, 학습 방법에 따른 인식 결과를 표 8에 나타내었다. 표 8에 나타난 것과 같이 손으로 분류하여 학습한 인식 결과가 인식기를 이용하여 음소를 분류한 후 학습시킨 결과보다 전반적으로 우수한 결과를 보였다. 그러나 표 8(a)에 나타난 결과 중 손으로 분류한 방법보다 자동 분류하여 학습시킨 인식 결과가 더 우수하게 나온 이유는 손으로 분류하기 어려운 음소간의 구분을 배열할 수 있기 때문이라고 사료된다.

표 8. 학습 방법에 따른 인식 결과(%)
Table 8. Recognition result(%) according to Alignment Methods

(a) 동일 데이터 사용

(a) Recognition Result using Training Speakers' Data

	남성 1	남성 2	여성 1	여성 2
자동 분류	93.2	81.2	93.2	70.8
손으로 분류	94.0	80.1	94.0	72.6

(b) 다른 데이터 사용

(b) Recognition Result using Non-Training Speakers' Data

	남성 3	남성 4	남성 5	남성 6	여성 3
자동 분류	90.2	87.3	79.8	88.6	75.4
손으로 분류	93.2	89.2	82.4	90.2	78.6

Ⅵ. 결론

본 논문은 한국어 음성 인식 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 음성 인식 결과와 학습시 학습 데이터의 분류 방법에 따른 인식 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 5개의 인식 단위를 추가하여 총 24개의 인식 단위를 설정하였다. 또한 자음의 위치 변화에 의한 음소의 수를 28개, 34개를 구분하였다. 따라서 최종 수행된 인식 단위의 총 수는 41개, 45개와 52개이다.

실험 결과, 모음의 경우 위와 같이 17개로 통합하여 인식 단위를 설정한 결과가 이론적 음소에 근거해 인식단위를 설정한 결과가 더 우수한 인식률을 나타내었다. 또한 자음의 위치 변화에 의한 음소 수 변화에 따른 인식 결과는 큰 변화를 보여주고 있지 않다. 이와 같은 결과는 학습시 음소의 발성에 따른 변화를 정확히 구분하지 않은 결과로 사료된다.

인식 방법에 따른 인식 결과는, 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. 분류 방법에 따른 인식 결과는 손으로 구분한 결과가 자동 분류 방법에 의한 결과보다 약간의 인식률 향상을 보였다.

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.

접수일자 : 2003. 1. 22 수정완료 : 2003. 3. 18

본 연구는 한국과학재단 (과제번호 R05-2002-000-00272-0)의 연구지원에 의해 수행된 것입니다.

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.

[9] Y. S. Hwang, "A Study on Korean Recognition Units for Speech Recognition System," proceeding of ICSP 2001, pp.375-378, 2001.

[10] 채나영, 황영수,"음소 분류에 따른 화자 적응 변화에 대한 연구," 2002한국신호처리.시스템학회추계학술대회논문집, pp.185-188, 2002.

황 영 수(YoungSoo Hwang)
正會員
1982 연세대학교 전자공학과 공학사
1984 연세대학교 전자공학과 공학석사
1990 연세대학교 전자공학과 공학박사
1989~현재 관동대학교 정보기술공학부 교수.
관심분야 : 음성신호처리, 음향공학 등