음성인식 기능을 가진 주소입력 시스템의 개발과 평가

Development and Evaluation of an Address Input System Employing Speech Recognition

김 득 수*, 황 철 준*, 정 현 열**
(Deok Soo Kim*, Cheol Jun Hwang*, Hyun Yeol Chung**)

요 약

본 논문은 음성인식 기술을 사용자 인터페이스로 하여 국내 행정 단위 시(도), 구(군), 동(읍,면), 번지로 구성되는 주소를 인식의 대상으로 하는 주소 입력 시스템 구축에 대하여 기술한다. 본 시스템은 사운드카드가 장착된 개인용 컴퓨터상의 윈도우 95환경에서 동작하며, 음성인식부는 인식의 기본단위로 유사음소단위(Phoneme Like Units; PLUs)를 이용하여 CHMM(Continuous Hidden Markov Model) 음소모델을 작성하고, 주소인식을 위해서 주소명의 특징을 고려하여 이에 적합한 유한상태 오토마타(Finite State Automata)를 구성하여 OPDP(One Pass Dynamic Programming)법으로 인식을 수행하였다. 실용성있는 시스템 성능을 얻기 위하여 마이크, 환경잡음 및 화자의 변화 등의 사용환경변화에 대해 최대사후확률 추정법(Maximum A Posteriori Probability Estimation; MAP)으로 적응화시켜 인식률의 향상을 도모하였고, 개인용 컴퓨터상에서의 인식속도를 향상시키기 위하여 가변프루닝 문턱치를 이용한 고속화 기법을 제안하였다. 평가결과, 화자적응화 후의 성인 남자 3인에 대한 100개의 연결주소명의 연결단어 인식률은 평균 96.0%이상, 인식속도는 발성완료후 약 2초 이내로 인식이 완료되어 본 시스템의 유효성을 확인할 수 있었다.

ABSTRACT

This paper describes the development and evaluation of a Korean address input system employing automatic speech recognition technique as user interface for input Korean address. Address consists of cities, provinces and counties. The system works on a window 95 environment of personal computer with built-in soundcard. In the speech recognition part, the Continuous density Hidden Markov Model(CHMM) for making phoneme like units(PLUs) and One Pass Dynamic Programming(OPDP) algorithm is used for recognition. For address recognition, Finite State Automata(FSA) suitable for Korean address structure is constructed. To achieve an acceptable performance against the variation of speakers, microphones, and environmental noises, Maximum a posteriori(MAP) estimation is implemented in adaptation. And to improve the recognition speed, fast search method using variable pruning threshold is newly proposed. In the evaluation tests conducted for the 100 connected words uttered by 3 males the system showed above average 96.0% of recognition accuracy for connected words after adaption and recognition speed within 2 seconds, showing the effectiveness of the system.

I. 서 론

최근 각종 미디어의 발달, 초고속 정보 통신망의 구축과 더불어 멀티미디어 통신을 통한 통신 판매, 고객 관리, 물류 처리, 제품 홍보 등이 폭증하고 있다. 이와 더불어 개인용 컴퓨터의 보급이 가속화됨에 따라 컴퓨터 시스템에 주소를 입력하여 고객을 관리하고, 상품을 배송하는 일이 보편화되고 있다. 이와 같은 단순 반복 수작업에 의한 자료의 입출력 및 검색은 음성인식 기능을 가진 컴퓨터를 이용하게 되면 신속하고 효율적인 처리가 가능하다.

음성인식 기능을 갖는 주소입력 시스템을 개발하는 데 있어 가장 핵심이 되는 음성인식 기술은 최근 반도체 메모리와 컴퓨터의 처리능력의 급속한 발전과 더불어 음성인식 기술도 괄목할 만한 발전을 가져와 연속음성인식에 있어서도 문장인식률 80%를 상회하고 있다. 이 기술을 이용하여 외국에서는 자동통역 시스템, 여행정보안내 시스템, 관광안내 시스템, 증권정보안내 시스템을 개발하

* 대구공업대학 전자계산과
** 영남대학교 정보통신공학과
접수일자 : 1998년 4월 3일

여 상품화하고 있으며 국내에서도 음성구동 퍼스널 컴퓨터, 증권정보안내 시스템이 개발되어 사용되고 있고, 미국, 일본 등과 같이 자동통역시스템 개발사업에도 참여하고 있다. 또 음성다이얼링 휴대폰을 개발하여 상용화중에 있다.

음성인식 기술을 이용한 주소입력 시스템 구축의 실용화 기술의 기본이 되는 기술로서는 지난 20여년 전부터 현재까지 계속되어온 연구결과로 자연언어(Spontaneous Speech) 인식으로 대표적인 대어휘 불특정화자 연속음성 인식 시스템으로는 미국의 카네기멜론대학의 SPHINX-II[1], JANUS-III[2], 스탠포드대학의 DRAGON Dictate, MIT의 SUMMIT, GALAXY, 일본의 ATR음성번역통신 연구소의 ASURA, 토요하시과학기술대학의 SPOJUS-SYNO 시스템[3]등이 있으며 현재 자연발성언어에 의한 연속음성에 대해 문인식률 80% 이상을 상회하고 있다. 또 이 기술을 이용하여 음성구동 개인용 컴퓨터(APPLE-MAC), 여행안내 시스템(MIT), 관광안내 시스템(Toyohashi University of Technology), 자동판매기(TOSHIBA)등을 개발하고 있다.

국내에서도 1980년도에 들면서부터 본격적인 음성인식에 관한 연구가 이루어져 오고 있으며 현재 외국의 예에서와 같이 한국전자통신 연구소의 자동통역시스템, 한국통신의 증권정보안내 시스템 등이 개발되어 있으며 음성구동 셀룰러폰도 현재 개발되어 상용되고 있다.

음성입력을 이용한 사용자 인터페이스에 관한 연구의 예를 들면 Pausch등은 컴퓨터상의 작도소프트웨어에 19종류의 명령어를 음성입력할 수 있는 기능을 부가하였다. 또 Shida등은 작도소프트웨어에 약 90단어의 음성입력 기능을 부가한 시스템을 구현하였다[4]. 이 경우 명령어 입력은 희망후보를 선택하는 것을 음성으로 대신함으로 인해 마우스의 이동량을 감소시켜 능률을 향상시키고 있다. 그렇지만 후보가 많을 경우 메뉴상에 많은 후보를 나타내어 그 중 희망하는 하나를 선택하는 것은 곤란하다. 따라서 음성입력에서는 희망하는 후보를 발성하면 입력될 수 있기 때문에 입력에 필요한 시간이 단축될 수 있을 뿐만 아니라 작업효율도 향상시킬 수 있다. 그러나 현재까지 개발된 많은 시스템의 경우, 대부분 특별한 하드웨어를 요구한다든지, 워크스테이션상에서 구동되는 경우가 대부분으로 실제로 일반인이 사용하는 데는 문제가 있었다.

따라서 본 연구에서는 범용 사운드카드가 장착된 개인용 컴퓨터의 윈도우 95환경에서 동작하는 음성인식 기능을 가진 주소입력검색 시스템을 구현하기로 하였다. 음성에 의한 주소입력의 경우, 오인식에 의한 입력오류가 발생할 수 있기 때문에 이를 보완하기 위해 마우스, 키보드, 터치스크린 기능을 부가하여 멀티모드(Multi-mode)화 하였다[5].

이하, II장에서는 저자들이 구현한 음성에 의한 주소입력 시스템의 개요, III장에서는 본 시스템에서 이용하고 있는 데이터베이스와 그 분석방법, IV장에서는 음성인식에 이용한 HMM학습과 적응화, V장에서는 주소 인식을 위한 연결단어 인식 방법, VI장에서는 기존의 고속 탐색 기법과 본 논문에서 제안하는 가변 프루닝 문턱치를 이용한 음성인식의 고속화에 대하여 설명하고, VII장에서는 인식실험 및 결과, 마지막으로 VIII장에서는 결론을 맺는다.

II. 시스템 개요

전체 시스템의 개략을 그림 1에, 화면구성도를 그림 2에 보인다. 본 시스템은 사운드카드를 내장한 개인용 컴퓨터상에서 동작하고 국내의 행정 단위 단어를 인식의 대상으로 한다. 주소음성인식에 있어서는 광역 단위로부터 Top-down방법으로 인식, 검색된다. 입력은 음성과 컴퓨터의 기본 입력 장치인 키보드, 마우스 모두를 통해서도 가능하다. 주소입력에 있어서는 최초 광역 단위인 전국 시도의 후보 단어를 윈도우 화면상에 나타내고 인식결과에 따른 하위 행정 단위의 후보만을 화면에 보여줌으로써 사용자가 주소를 선택 하는 데 있어서의 오류를 방지한다.

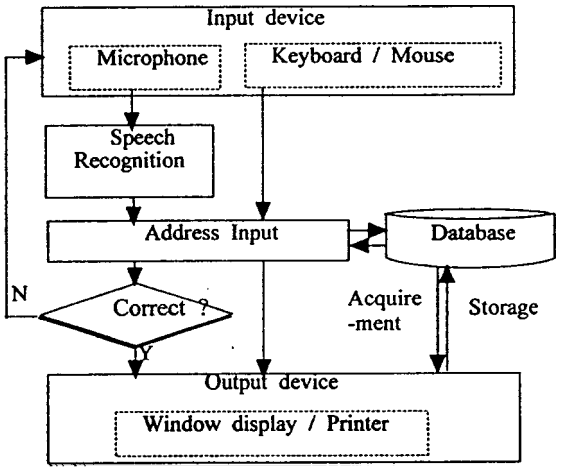


그림 1. 전체 시스템의 개략도
Fig. 1. Overview of the Address input system.

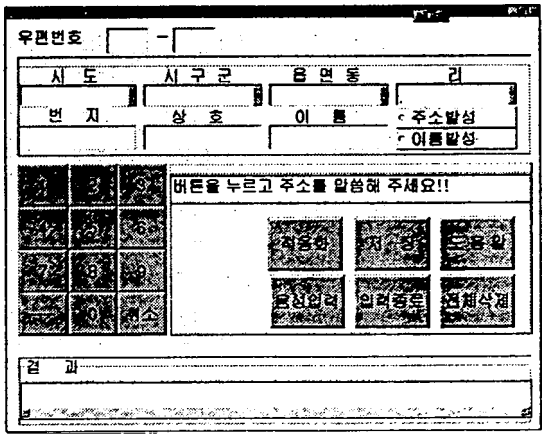


그림 2. 음성인식 주소입력 시스템의 화면구성
Fig. 2. Window for the Address input system.

III. 음성 데이터 및 분석

화자독립 기본모델(Speaker Independent HMM; SI-HMM)은 한국전자통신 연구원(ETRI)에서 작성한 PBW-(Phoneme Balanced Words) 445단어 음성 데이터베이스(이하 ETRI445)중 14인의 1회 발성을 이용하여 구성한다. 적응화와 인식단계에 있어서는 사무실환경에서 3인의 남성화자가 2개의 서로 다른 마이크를 이용하여 발성한 600개 연결단어음성을 이용한다. 표 1은 실험에 사용된 음성 데이터베이스를 나타낸다.

발성된 주소음성 데이터는 16kHz/16bits로 A/D변환된 후, 에너지와 영교차율의 평균과 분산을 이용하여 음성 구간을 검출한다. 검출된 주소음성으로부터 14차의 LPC 분석을 통하여 10차의 멜켑스트럼 계수(Mel Frequency Cepstral Coefficient; MFCC)를 구하고, 이 멜켑스트럼 계수와 그 회귀계수(Regressive Coefficient; RGC)를 음성특징 파라미터로 한다. 인식실험시 10차의 MFCC와 10차의 RGC를 사용한다. 표 2에 특징 추출을 위한 음성 자료의 분석 조건을 나타낸다.

표 1. 음성 데이터베이스
Table 1. Speech data.

Speakers (Number)	Male(14)	Male(3)	
Usage Step	Training	Adaptation	Test
# of Utterances	445 words	25 connected words	75 connected words
Recording Device	DAT	PC (Sound card)	
Environ -ment	Sound proof booth	office	
Micro -phone	Dynamic, Headset	Condenser desktop Dynamic Headset	

표 2. 음성 데이터의 분석조건
Table 2. Analysis condition of speech data.

Sampling Frequency	16 kHz
Resolution	16 bits
Hamming Window	16 msec (256 points)
Frame Rate	5 msec (80 points)
Analysis	14 order LPC
Feature Parameters	10 order MFCC+10 order RGC

IV. HMM 학습과 적응화

그림 3에 본 시스템의 학습 및 평가의 전 과정을 개략적으로 나타낸다. 주소음성인식을 위한 인식의 기본단위는 48개의 유사음소로 하고, 각 유사음소의 음향모델에 사용한 HMM은 4상태 3출력분포의 연속출력분포형 HMM(이산분포형 지속시간 제어)을 사용한다. 마이크의 변동, 사용환경변화에 대한 인식률 제고를 위하여 최대사후확률분포를 이용한 적응화 기법[6]을 이용하여 음소 HMM모델(SI-HMM)을 적응화하고, 인식단계에서는 적응화된 음소모델을 이용하여 OPDP알고리즘으로 인식한다. 적응화와 인식은 모두 사무실 환경하에서 이루어진다.

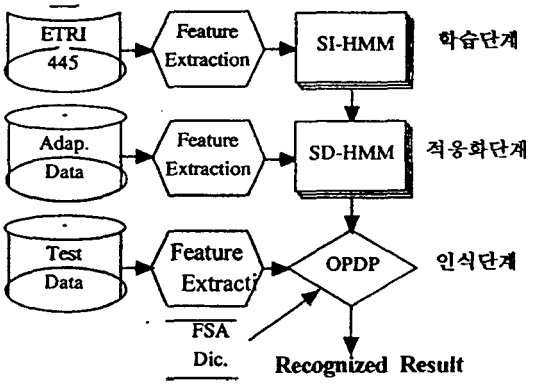


그림 3. 학습 및 평가의 흐름도
Fig. 3. Block diagram for training and tests.

표 3. 마이크변화에 따른 평균 단어 인식률
Table 3. Average recognition rate for different microphone.

Speaker	mj(%)		ks(%)		cj(%)	
Mic.	Desktop	Headset	Desktop	Headset	Desktop	Headset
CWRR	86.7	88.0	92.0	84.0	85.3	88.0

- CWRR : Connected Words Recognition Rate

일반적으로 학습용 데이터와 평가용 데이터의 사이에 마이크를 포함한 녹음 환경의 차이, 화자 변화에 따른 발성의 차이 등은 인식률에 직접적인 영향을 미친다. 이를 해소하기 위해 그림 3에서 보인 바와 같이 적응화 학습을 실시한다. 기존의 HMM을 이용한 음성인식기의 대부분은 ML(Maximum Likelihood)에 기반을 둔 Baum-Welch 학습법으로 파라미터를 재추정하고 있다. ML학습은 기본적으로 무한한 양의 학습데이터와 각 모델이 서로 독립적이라는 가정을 기초로 한다. 그러나 실제적인 학습의 경우 각 모델들은 서로 독립적으로 보기 어렵고 학습데이터 양도 상당히 제약되어 있어서 인식기의 변별력을 저하시키는 원인이 된다. 반면에 최대사후확률추정법을 이용하면 적응화가 중단되어도 그 시점까지 최적인 파라미터를 추정할 수 있고 필요시 추가적으로 적응화를 수행해 파라미터의 정밀도를 향상시킬 수 있다. 사후확률 추정식은 식 (1)과 같이 나타낸다. 식 (1)을 이용하면 1개의 학습샘플이 주어진 경우에도 사후확률이 최대가 되도록 파라미터를 추정할 수 있다

$$\max_{\theta} P(\theta \mid X_1, \ldots, X_N)$$

$$= \max_{\theta} \frac{P(X_n \mid X_1, \ldots, X_{N-1}, \theta)\, P(\theta \mid X_1, \ldots, X_{N-1})}{\int P(X_n \mid X_1, \ldots, X_{N-1}, \theta)\, P(\theta \mid X_1, \ldots, X_{N-1}) d\theta} \quad (1)$$

최대사후확률추정법을 이용해 정규분포의 평균과 분산을 동시에 추정하고자 하는 경우에는 파라미터가 평균(μ)과 분산($\sum$)이 된다[7]. 이때 N개의 학습샘플을 이용한 평균벡터의 재추정식은 다음 식과 같이 표시할 수

$$\hat{\mu}_N = \frac{\alpha\mu_0 + \sum_{i=1}^{N} X_i}{\alpha + N} \tag{2}$$

또한 공분산 행렬의 재추정식은

$$\sum_N = \frac{1}{\beta+N}\{X_N X_N^T - (\alpha+N)\mu_N\mu_N^T + (\beta+N-1)\sum_{N-1} + (\alpha+N-1)\mu_{N-1}\mu_{N-1}^T\} \tag{3}$$

와 같다. 식 (2)와 식 (3)에서 α, β는 초기 HMM의 평균과 공분산 행렬을 추정하기 위해 사용한 샘플수이다. 이 값은 실험에 의해 재조정할 필요는 있지만 인식률에는 그다지 영향을 끼치지 않는 것으로 알려져 있다[8]. 본 연구에서는 전 화자에 대해 적응화 계수 $\alpha=15$, $\beta=50$으로 일정하게 부여한다.

V. 연결단어 인식

주소입력 시스템에서 인식의 대상이 되는 주소의 경우 고립단어가 여러 개 나열된 연결단어 형태로 구성되어 있기 때문에 고립단어를 대상으로 하는 인식방법으로는 제약이 따르게 된다. 만약 고립단어인식법으로 연결 주소명을 인식하는 경우, 한 단어를 발성하고 이를 인식하여 결과를 확인하고 또 다시 하위 단위의 단어를 발성하여 인식을 수행해야 하기 때문에 사용상 불편이 크다.

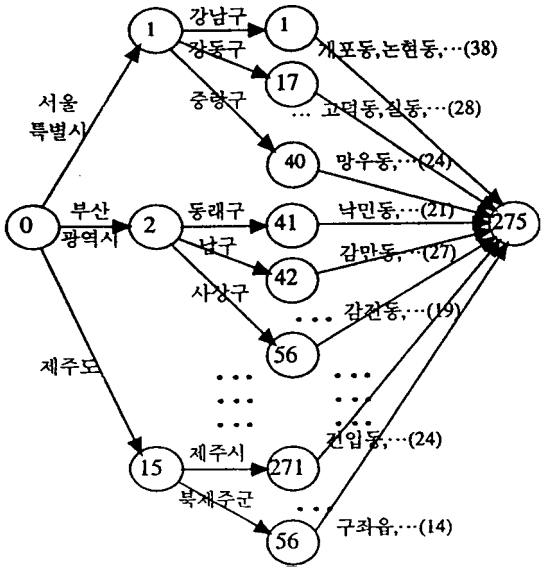


그림 4. 주소명인식을 위한 유한상태 오토마타
Fig. 4. Finite State Automata for Korean address.

따라서 주소의 경우 연결단어 즉, "대구광역시 수성구 만촌동" 등과 같은 주소명 특유의 연속적인 발성으로 확장하여야 한다. 이를 위하여 본 논문에서는 주소입력 시스템의 인식대상이 주소라는 점을 고려한 연결단어를 위한 유한상태 오토마타(Finite State Automata)를 구성한다. 고립단어인식에서 연결단어로 인식 대상을 확장하기 위한 전국 주소명에 대한 유한상태 오토마타를 그림 4에 나타낸다. 그림에서 나타낸 것과 같이 처음 시작상태에서는 전국 광역시도 단위를 대상으로 인식을 수행하고, 인식된 시도 단위에 해당하는 하위 행정 단위를 인식하며 같은 방법으로 그 하위 행정 단위를 인식하도록 계층적인 구조로 오토마타를 구성한다.

VI. 가변 프루닝 문턱치를 이용한 음성인식의 고속화

인식에 있어서 가장 간단한 방법은 예측되어진 전체의 후보와 입력음성을 정합시키는 방법이다. 그러나 이 방법은 대상 어휘수가 증가하고 인식 알고리즘이 복잡해짐에 따른 대규모 탐색공간이 필요한 경우에 대해서는 많은 처리시간을 요구한다. 실시간 음성인식을 위해서는 전체의 후보와 정합을 행하지 않고도 고정도의 인식성능을 얻을 수 있는 효율적인 탐색수법이 필요하게 된다.

이하에 탐색 공간을 줄이기 위한 방법으로 고속 탐색기법과 목구조사전 구성에 대해 기술하고, 본 논문에서 제안하는 프레임동기형 가변프루닝 문턱치를 이용한 탐색공간 감소법에 대해 설명한다.

6.1 고속 탐색기법

음성인식을 수행하기 위해서는 출력확률 계산과 탐색의 2가지 계산과정을 필요로 한다. HMM을 이용한 음성인식에서의 출력확률 계산은 임의의 한 시점에서 관측된 음성을 출력하는 주어진 HMM의 상태의 확률계산이며, 탐색은 주어진 음성 입력에 대한 최상의 상태열을 구하는 문제로 볼 수 있다. 이러한 탐색에 소요되는 시간은 음향학적 모델의 복잡성에 의해서는 크게 영향을 받지 않으나, 인식대상의 규모에 따른 영향은 크다. 즉, 인식에 있어서 모든 가능한 상태열들을 고려할 경우, 입력된 음성에 대한 최고 우도의 상태열(단어, 문장)을 찾기 위한 탐색공간은 지수함수적으로 증가한다.

본 논문에서는 탐색공간제한에 있어서 효율적인 빔 탐색법을 도입한다. 빔 탐색법[9]은 연속음성인식의 탐색법으로써 단어수 동기형으로 1974년에 최초로 발표되었는데 이 수법은 일단 인식되어진 음운 계열에 대해서 단어열을 탐색하는 방법이다. 1976년에 발표되어진 HARPY 연속음성인식시스템에서는 프레임동기형 네트워크기반의 인식모델로써 우도의 낮은 상태를 프루닝하는 수법을 이용한 빔 탐색법을 채용하였다[10].

현재까지 대부분의 시스템에서는 프레임동기형의 빔 탐색법을 이용하고 있는데 이 방법은 각 후보의 우도를 비교하고 상위 일정 개수(문턱치이하의 것)에 대해서만 후속 정합을 고려하는 방법으로 다음과 같이 나타낸다[11].

$$D_{\min}(i,j) \le D_{\min}(i,j^*) + \delta \quad (4)$$

이 방법은 i프레임에서의 최적의 경로 (i,j^*)에 대해 문턱치 δ이내의 상위 몇 개(빔 폭)만을 후속탐색에서 고려하고 나머지는 탐색으로부터 제외하는 방법이다. 정합은 입력 프레임과 식 (4)의 범위내의 노드에 대응하는 음향 모델과의 정합을 의미한다. 여기서 각 노드의 우도를 비교하여 상위 일정 개수를 선택한 후, 여기서부터 전개되어지는 노드들과 입력 $i+1$프레임과 정합한다.

탐색공간을 더욱 제한하는 방법으로써 프루닝 기법이 있다. 이 방법은 각 프레임에 있어서 최대 우도를 $g_{\max}$로 하고, $g_{\max} = -\lambda$ (λ는 여유분을 둔 문턱치) 에 만족하지 않는 후보에 대해서는 그 시점 이후의 탐색을 프루닝함으로써 탐색공간을 감소시킨다.

먼저 One-pass Viterbi 알고리즘의 누적대수우도확률 $p_q^n(i,j)$의 i프레임에 대한 최대치 $P_{\max}(i)$를 다음과 같이 구한다.

$$P_{\max}(i) = \max_{j,n,q} P_q^n(i,j) \quad (5)$$

이렇게 구해진 최대우도에 대해 식 (6)과 같은 조건을 만족하는 각 상태 q의 각 단어(또는 PLU)에 대해서만 탐색을 수행하고 나머지는 제외하는 기법을 다음 식으로 나타낼 수 있다.

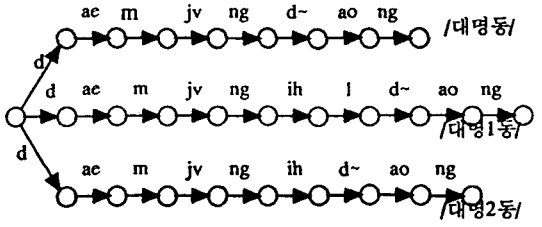
$$\max_j P_q^n(i,j) < P_{\max}(i) - \lambda \quad (6)$$

빔 탐색법에서 가장 중요한 것은 각 후보의 우도의 정도이다. 정도가 낮은 경우, 정해로 얻어진 후보가 프루닝에 의해 제외되는 오류가 있을 수 있다. 즉, 어떤 시점(처리 프레임)에서 그 노드까지의 누적우도가 크지 않을 경우 정해가 될 수 있음에도 불구하고 탐색에서 제외되어 최적성을 보장받지 못하게 되므로, 빔 폭의 제한과 프루닝조건을 엄격하게 함으로써 최적해를 잃을 우려가 있다. 따라서, 인식정도에 영향을 주지 않기 위해서는 빔 폭과 프루닝조건을 완화시키면서 탐색공간을 감소시키는 방법을 찾을 필요가 있다.

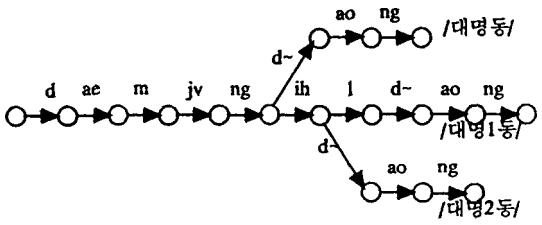
6.2 목구조형 사전

일반적으로 탐색되어지는 상태 네트워크(음소열)가 어휘내의 각 단어에 대해 선형적으로 결합되어 있기 때문에, 이러한 구성으로 인해 실제로 탐색되는 모델의 수는 어휘수에 비례하게 된다.

따라서, 탐색공간의 크기를 줄이기 위해서는 목구조형 사전(Tree-structured lexicon)을 구성하는 방법을 도입할 필요가 있다[12][13]. 이것은 고속화를 위한 언어적 수법 중의 하나로써 많은 단어들이 어두부분에 동일한 접두어로 시작하고 있는 점에 착안하여 이를 공유하여 중복을 피하므로서 탐색공간을 줄이는 방법이다. 초기에 목구조는 여러 시스템에서 후보성분을 생성하기 위한 고속 정합에 사용되었다. 그림 5에서는 선형결합과 목구조형 사전구성의 예를 나타낸다.



(a) Linearly connected lexicon.



(b) Tree structured lexicon.

그림 5. PLUs를 이용한 한국어 단어사전의 예
Fig. 5. Examples of lexicon for Korean word based on PLUs.

6.3 제안된 고속화 방법

6.1절에서 서술한 바와 같이 오토마타제어에 기초한 연속음성인식 알고리즘은 프레임에 동기하여 인식처리의 탐색공간을 확장하는 비교적 효율적인 방법이다. 그러나, 프레임의 수가 증가함에 따라 후속하는 오토마타의 상태수가 증가하게 되면, 계산량이 지수함수적으로 증가하게 되므로 대어휘연속음성인식에 있어서는 치명적인 문제를 초래하게 된다. 이러한 계산량증가의 문제를 해결하기 위해 6.2절에서의 목구조의 사전정보를 이용한 언어적 수법을 도입하고, 확률에 대한 제한을 통한 탐색공간의 감소법을 도입하면 대어휘에서도 효율적인 처리가 가능할 것으로 생각된다.

여기서는 프루닝 기법에 의한 탐색공간의 제한에 있어서 프레임의 진행에 따라 구해지는 누적확률에 근거하여 문턱치의 설정을 가변시키는 프레임동기형 가변 프루닝문턱치의 도입을 제안한다. 이는 OPDP(One-Pass Dynamic Programming)법의 특성인 프레임 동기성의 장점을 이용하고, 최적해를 위한 누적확률이 프레임의 진행에 따라 분포의 폭이 작아지며 수렴화하는 것에 기초한 것이다. 전술한 식 (6)에 대하여 본 논문에서 제안하는 가변 프루닝문턱치를 설정할 경우 다음과 같이 표현된다.

$$\max_j P_q^n(i,j) < P_{\max}(i) - \lambda(k) \quad (7)$$

식 (7)에서 문턱치 λ를 설정하는 데 있어서 프레임 i에서의 정수값 k에 따른 가변 프루닝 문턱치 $\lambda(k)$를 도입하고 있다. 정수값 k의 설정은 프레임 i에 대해 선형적 ($k=ai,\ a=1,2,3,\cdots$ or $1/2,1/3,\cdots$) 또는 계단식 ($k=10,20,30,\cdots$)으로 증가량을 설정할 수 있다. 그림 6에 기존의 고정상수형 프루닝문턱치 λ의 적용할 경우와 본 논문에서 제안하는 프레임동기형 가변프루닝 문턱치 $\lambda(k)$를 적용할 경우에 대한 탐색공간을 비교한 예를 보인다.

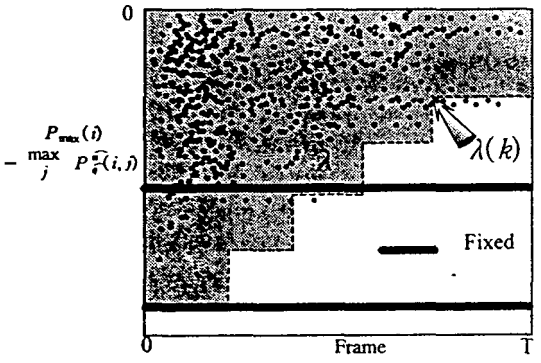


그림 6. 고정상수형 및 프레임동기형 가변프루닝 문턱치
Fig. 6. Fixed and frame-synchronous variable pruning threshold.

그림에서 누적대수우도확률과 $i \to j$ 프레임까지 진행되는 시점에서의 우도확률과의 차를 시간경과에 따라 나타낸 것으로 프레임 종단으로 갈수록 분포의 폭이 좁아짐을 알 수 있다. 이러한 가변 프루닝 문턱치를 사용함으로써 탐색의 공간을 더욱 제한함은 물론 2.1절에서 언급한 바와 같이 상수값으로 지정할 경우 일어날 수 있는 최적해를 잃지 않는 효과를 기대할 수 있다.

VII. 인식 실험 결과

본 시스템의 인식 알고리즘은 OPDP법을 사용한다. 그리고 사용된 음성데이터는 ETRI 445 단어 데이터베이스와는 달리 조용한 사무실 환경하에서 100개의 연결주소명을 남성화자 3명이 2종의 서로 다른 마이크를 사용하여 채록한 연결주소 데이터를 이용하였다. 총 600개 연결단어 (100 연결단어 × 3인 × 2 마이크 × 1회)이며, 목구조는 한국어 행정단위에 따라 서로 다른 어휘수로 구성되었다.

표 4와 표 5에 적응화와 인식실험에 사용된 한국어 행정 주소명 데이터베이스를 나타낸다.

실제 음성인식 시스템에 사용되는 여러 환경에 대하여 고정도 인식률을 얻기 위하여 앞에서 설명한 최대사후확률추정법에 의한 적응화 기법을 도입하여 초기 화자독립 HMM을 적응화한다. 적응화를 위해서는 대구광역시 연결주소명 100개중 25개를 이용하고 인식실험에는 적응화에 참여하지 않은 75개를 이용하였다.

앞에서 설명한 고속화 기법중 고정 프루닝 문턱치를 이용하여 인식실험을 실시하였다. 이때 HMM 모델은 적응화 후의 모델을 사용하고, 빔 폭은 10으로 하여 프루닝 문턱치의 값만을 변화시키면서 인식실험을 수행하였다. 이 결과를 표 6에 나타내었다.

표 4. 적응화용 데이터베이스
Table 4. Database for adaptation.

번호	Connected Words	번호	Connected Words
1	대구광역시 남구 대명3동	14	대구광역시 수성구 가천동
2	대구광역시 남구 이천동	15	대구광역시 수성구 만촌3동
3	대구광역시 달서구 두류동	16	대구광역시 달성군 유가면
4	대구광역시 달서구 용산동	17	대구광역시 달성군 하빈면
5	대구광역시 동구 능성동	18	대구광역시 중구 남산1동
6	대구광역시 동구 사복동	19	대구광역시 중구 대신1동
7	대구광역시 동구 신암1동	20	대구광역시 중구 도원동
8	대구광역시 동구 용수동	21	대구광역시 중구 동성로1가
9	대구광역시 북구 노원1가	22	대구광역시 중구 서양동
10	대구광역시 북구 대현1동	23	대구광역시 중구 인교동
11	대구광역시 북구 복현동	24	대구광역시 중구 태평로2가
12	대구광역시 북구 서변동	25	대구광역시 중구 하서동
13	대구광역시 북구 읍내동		

프루닝 문턱치가 400인 경우, 인식률의 저하없이 기본 탐색에 비하여 인식시간이 약 1/100로 감소하여 이 알고리즘의 유효성을 확인할 수 있다. 이를 개인용 컴퓨터 환경하에서 동작시켰을 경우 별도의 하드웨어 추가없이 연결단어의 평균 발성시간이 약 2.5초 정도가 소요될 경우, 발성이 끝나는 시점으로부터 약 2초내에 인식이 완료함을 확인할 수 있어 실시간 처리가 가능함도 알 수 있었다.

다음은 본 논문에서 제안하는 빔 탐색법의 문제점을 고려하여 프레임이 진행됨에 따라 가변되는 프레임동기형 프루닝 문턱치를 이용하여 인식실험을 수행하였다. 이때의 인식실험 결과를 표 7에 나타내었다.

프레임동기형 가변프루닝 문턱치를 사용하였을 경우 표에서 보는 바와 같이 인식률이 저하되지 않으면서 콘덴서 데스크탑 마이크의 경우 0.16초, 다이나믹 헤드셋 마이크의 경우 0.65초의 인식시간이 단축되어 제안된 고속화 알고리즘의 유효성을 확인할 수 있었다.

VIII. 결 론

본 논문에서는 음성인식 기능을 가진 주소 입력시스템을 구축하여 한국어 행정 주소명 100개에 대한 인식실험 결과 시스템의 유효성을 확인하였다.

마이크의 변화와 같은 발성환경 변화요인에 의한 인식성능 저하를 확인하기 위하여 대구광역시 연결단어 75개를 대상으로 적응화전 사전 인식실험을 수행한 결과 화자의 변화에 대해서는 1.3~6.7%, 마이크의 변화에 대해서는 1.3%~8.0%의 인식률 변화가 있음을 확인하였다.

실용화 가능한 시스템의 성능을 얻기위하여 마이크, 환경잡음 및 화자의 변화 등의 사용환경변화에 대해 최대사후확률추정법(식성능 저하를 해결하기 위해 최대사후

표 5. 테스트용 데이터베이스
Table 5. Database for test.

번호	Connected Words	번호	Connected Words
1	대구광역시 남구 대명동	39	대구광역시 서구 원대2가
2	대구광역시 남구 대명6동	40	대구광역시 서구 중리동
3	대구광역시 남구 봉덕동	41	대구광역시 서구 평리3동
4	대구광역시 남구 봉덕2동	42	대구광역시 서구 평리6동
5	대구광역시 달서구 갈산동	43	대구광역시 수성구 고모동
6	대구광역시 달서구 대곡동	44	대구광역시 수성구 내환동
7	대구광역시 달서구 도원동	45	대구광역시 수성구 두산동
8	대구광역시 달서구 두류3동	46	대구광역시 수성구 삼덕동
9	대구광역시 달서구 송현동	47	대구광역시 수성구 성동
10	대구광역시 달서구 유천동	48	대구광역시 수성구 시지동
11	대구광역시 달서구 죽전동	49	대구광역시 수성구 욱수동
12	대구광역시 달서구 파호동	50	대구광역시 수성구 지산동
13	대구광역시 동구 각산동	51	대구광역시 수성구 지산2동
14	대구광역시 동구 검사동	52	대구광역시 달성군 구지면
15	대구광역시 동구 덕곡동	53	대구광역시 달성군 다사면
16	대구광역시 동구 도학동	54	대구광역시 중구 공평동
17	대구광역시 동구 둔산동	55	대구광역시 중구 남산동
18	대구광역시 동구 방촌동	56	대구광역시 중구 남산4동
19	대구광역시 동구 봉무동	57	대구광역시 중구 남성로
20	대구광역시 동구 부동	58	대구광역시 중구 대봉동
21	대구광역시 동구 송정동	59	대구광역시 중구 덕산동
22	대구광역시 동구 신암4동	60	대구광역시 중구 동문동
23	대구광역시 동구 신평동	61	대구광역시 중구 동산동
24	대구광역시 동구 율암동	62	대구광역시 중구 동성로3가
25	대구광역시 동구 진인동	63	대구광역시 중구 동인2가
26	대구광역시 북구 고성2가	64	대구광역시 중구 동인4가
27	대구광역시 북구 금호동	65	대구광역시 중구 북성로2가
28	대구광역시 북구 동변동	66	대구광역시 중구 사일동
29	대구광역시 북구 동호동	67	대구광역시 중구 상서동
30	대구광역시 북구 사수동	68	대구광역시 중구 서문로2가
31	대구광역시 북구 산격동	69	대구광역시 중구 서성로2가
32	대구광역시 북구 칠성1가	70	대구광역시 중구 수창동
33	대구광역시 북구 침산3동	71	대구광역시 중구 시장북로
34	대구광역시 북구 학정동	72	대구광역시 중구 용덕동
35	대구광역시 서구 내당4동	73	대구광역시 중구 장관동
36	대구광역시 서구 비산3동	74	대구광역시 중구 종로2가
37	대구광역시 서구 비산5동	75	대구광역시 중구 태평로2가
38	대구광역시 서구 상리동		

표 6. 고정 프루닝 문턱치를 이용한 고속알고리즘을 이용한 인식실험 결과
Table 6. Recognition results by fast algorithm using fixed pruning threshold.

Microphone			Desktop		Headset	
	BW	Pr_th	CWRR (%)	Time (sec)	CWRR (%)	Time (sec)
Adap (×)	10	-500	88.0	30.49	86.7	28.59
Adap (○)		-500	96.0	6.05	96.0	5.43
		-400	96.0	5.26	96.0	4.65
		-300	95.6	4.24	95.6	4.06
		-200	94.7	3.03	95.6	3.4

확률추정법에 의한 적응화 기법을 도입하여 인식실험을 수행한 결과 적응화전 연결단어 평균 87.3%의 인식률에 비해, 적응화후 연결단어 평균 96.0%의 높은 인식률을 얻어 적응화 방법의 유효성을확인하였다.

개인용 컴퓨터상에서의 인식속도를 향상시키기 위하여 가변프루닝 문턱치를 이용한 고속화 기법을 제안하였다.

표 7. 가변 프루닝 문턱치를 이용한 인식실험 결과
Table 7. Recognition results by fast algorithm using variable pruning threshold.

Microphone		Desktop		Headset	
Beam Width	Pruning Threshold	CWRR (%)	Time (sec)	CWRR (%)	Time (sec)
10	FSV1	96.0	5.10	96.0	4.51
	FSV2	95.1	4.78	96.0	4.39
	FSV3	92.9	4.06	94.7	3.80
	FSV4	95.6	4.47	96.4	4.00
	FSV5	88.9	3.35	88.4	3.09

단, FSV1: 초기치=-500, 100프레임마다 50씩 감소
FSV2: 초기치=-500, 100프레임마다 60씩 감소
FSV3: 초기치=-500, 50프레임마다 50씩 감소
FSV4: 초기치=-400, 100프레임마다 50씩 감소
FSV5: 초기치=-400, 50프레임마다 50씩 감소

가변프루닝 문턱치를 이용하여 인식한 결과 기존의 고정 프루닝 문턱치를 이용한 경우보다 평균 약 0.4초의 인식시간의 감소를 보여 제안한 알고리즘의 유효성을 확인하였다.

전체 시스템의 평가결과, 화자적응화 후의 성인 남자 3인에 대한 100개의 연결주소명의 연결단어 인식률은 평균 96.0%이상, 인식속도는 발성완료후 약 2초 이내로 인식이 완료되어 본 시스템의 유효성을 확인할 수 있었다.

참 고 문 헌

1. Alleva, F., et al., "Applying SPHINX-II to the DARPA Wall Street Journal CSR task," Proc. of Speech and Natural Language Workshop, pp. 393-398, Feb. 1992.
2. Alon Lavie, et al, "JANUS-III: Speech-to-speech translation in multiple languages," Proc. IEEE ICASSP-97, Vol.1, pp. 99-102, April 1997.
3. A. kai and S. Nakagawa, "A frame-synchronous continuous speech recognition algorithm using a top-down parsing of context-free grammar," Proc. ICSLP 92, pp. 257-260, 1992.
4. T. Nishimoto, N. Shida, T. Kobayashi, K. Shirai, "Multimodal Drawing Tool Using Speech, Mouse and Keyboard," Porc. ICSLP, Vol. 3, pp. 1287-1290, 1994.
5. Katsuhiko Shirai, "Spoken Dialogue in Multimodal Human Interface," ICSP '97, pp. 13-20, Aug.. 1997.
6. Rabiner and Juang, "Fundamentals of Speech Recognition," Prentice-Hall International, Inc, 1993.
7. Keinosuke Fukunage, "Introduction to Statistical Pattern Recognition," 2nd Edition, Academic Press, 1990.
8. Todashi koshigawa, "A Study on Speaker Adaptation of HMM in a Continuous Speech Recognition," Master thesis of Toyohashi, 1993.
9. 坂井利, 中川聖一, "構文情報を用いた連続音聲認識," 情報處理學會第15回全國大會, 37, Dec. 1994.
10. B. T. Lowerre, "HARPY speech recognition system," PhD thesis, Canegie-Mellon-University, 1976.
11. S. Furui, "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum," IEEE Trans. Acoustics, Speech and Signal Processing, Vol. 34, No. 1, pp. 52-59, Feb. 1986.

12. P. S. Gopalakrishnan, L. R. Bahl and R. L. Mercer, "A tree search strategy for large-vocabulary continuous speech recognition," Proc. IEEE ICASSP-95, Vol. 1, pp. 572-575, May 1995.
13. F. Richardson, M. Ostendorf and J. R. Rohlicek, "Lattice-based search strategies for large vocabulary speech recognition," Proc. IEEE ICASSP-95, Vol. 1, pp. 576-579, May 1995.

▲김 득 수(Deok-Soo Kim) 1956년 3월 3일생
1997년 2월 : 한양대학교 전자공학과 졸업(공학사)
1986년 2월 : 영남대학교 대학원 전자공학과 졸업
1990년 9월 ~ 현재 : 영남대학교 대학원 전자공학과 박사과정(수료)
1983년 3월 ~ 현재 : 대구공업대학 전자계산과 부교수
※주관심분야 : 음성분석 및 인식, 디지털신호처리

▲황 철 준(Cheol-Jun Hwang) 1970년 10월 24일생

1996년 2월 : 영남대학교 전자공학과 졸업(공학사)
1998년 2월 : 영남대학교 대학원 전자공학과 졸업(공학석사)
1998년 3월 ~ 현재 : 영남대학교 대학원 전자공학과 박사과정
※주관심분야 : 음성분석 및 인식, 디지털신호처리

▲정 현 열(Hyun-Yeol Chung)
현재 : 영남대학교 정보통신공학과 부교수
한국음향학회지 16권 8호(1997) 참조