

論文98-35C-11-8

# 은닉 마르코프 모델과 레벨 빌딩을 이용한 한국어 연속 음성 인식

## (Recognition of Continuous Spoken Korean Language using HMM and Level Building)

金 庚 炫 \* , 金 相 均 \*\* , 金 恒 俊 \*\*\*

(Kyung Hyun Kim, Sang Kyoon Kim, and Hang Joon Kim)

### 요 약

한국어 연속 음성에서 발생하는 조음결합문제를 해결하기 위하여 단어를 기본 인식 단위로 사용할 경우 각 단어의 효율적인 표현 방법, 연속된 단어로 이루어진 여러 문장의 표현 방법 그리고 입력된 연속음을 연속된 여러 단어로의 정합 방법에 관한 연구가 선행되어야 한다. 본 논문에서는 은닉 마르코프 모델과 레벨 빌딩 알고리즘을 이용한 한국어 연속 음성 인식 시스템을 제안한다. 각 단어는 은닉 마르코프 모델로 표현하고 문장을 표현하기 위하여 단어 모델을 연결한 형태인 인식 네트워크를 구성한다. 인식네트워크의 탐색 알고리즘으로는 레벨 빌딩 알고리즘을 사용한다. 제안한 방법은 항공기 예약 시스템에 적용한 실험에서 인식율과 인식속도면에서 실용적이었으며 또한 비교적 적은 저장공간으로 전체 문장을 표현하고 쉽게 확장할 수 있다는 장점을 가지고 있다.

### Abstract

Since many co-articulation problems are occurring in continuous spoken Korean language, several researches use words as a basic recognition unit. Though the word unit can solve this problem, it requires much memory and has difficulty fitting an input speech in a word list. In this paper, we propose an hidden Markov model(HMM) based recognition model that is an interconnection network of word HMMs for a syntax of sentences. To match suitably the input sentence into the continuous word list in the network, we use a level building search algorithm. This system represents the large sentence set with a relatively small memory and also has good extensibility. The experimental result of an airplane reservation system shows that it is proper method for a practical recognition system.

\* 正會員, 現代重工業 麻北里研究所

(Mabook-li Research Center, Hyundai Heavy Industries)

\*\* 正會員, 仁濟大學校 情報컴퓨터學部

(School of Information and Computer Inje Univ.)

\*\*\* 正會員, 慶北大學校 컴퓨터工學科

(Dept. of Computer Engineering Kyungpook Univ.)

※ 이 논문은 1997년 한국학술진흥재단의 공모과제 연구비에 의하여 연구되었음

接受日字:1998年6月2日, 수정완료일:1998年10月23日

### I. 서 론

음성 인식의 궁극적 목적은 인간과 컴퓨터사이의 보다 원활한 의사소통 수단을 제공하여 보다 편리한 인터페이스를 제공하는 것이다. 그러나 수년간의 많은 연구에도 불구하고 음성 인식은 다음과 같은 이유로 인하여 많은 어려움을 가지고 있다.

첫째로 화자독립의 문제이다. 화자에 따라 음성은 많은 변이를 가지므로 이러한 변이에 적절히 대응할 수 있는 방법이 요구된다. 둘째로는 보다 사람처럼 사람의 음성을 인식하기 위해서는 고립된 단어 하나의

인식보다는 연속적으로 발음되는 말을 인식하여야 한다. 세 번째는 보다 실용적이고 인간에 근접한 인식 시스템이 되기 위해서는 인식할 수 있는 말의 수가 많아야 한다. 마지막으로 문법적 제약이 없는 문장을 인식할 수 있어야 한다<sup>[1]</sup>.

위의 네 가지 어려움 중 인식 단어 수와 문법 구조가 제한될 경우 한국어 연속음성 인식은 주로 음절(syllables) 또는 음소(phonemes)와 같은 서브워드(subword)를 인식의 기본단위로 하여 연구되어 왔다<sup>[2-5]</sup>. 이러한 서브워드를 기본단위로 사용할 경우 서브워드 사이에서 발생하는 조음결합 문제(co-articulation problem)로 인하여 인식에 많은 어려움을 가지게 된다<sup>[6]</sup>. 따라서 제한된 언어를 사용하기 때문에 인식 단어의 수가 많지 않은 음성 예약 시스템, VDS(Voice Dialing Service)에서는 서브워드 대신 단어를 인식의 기본단위로 사용하여 단어 내부의 조음결합 문제를 해결하고 인식율을 향상시킬 수 있다<sup>[7]</sup>.

그러나 단어를 기본 인식단위로 사용할 경우 인식속도와 저장공간의 문제를 해결해야 한다. 즉 각 단어의 효율적 표현 방법과 연속된 단어로 이루어진 여러 문장의 표현 방법 그리고 입력된 연속음성을 연속된 여러 단어로의 정합 방법에 관한 연구가 필수적이다.

본 논문에서는 이러한 단어단위의 연속 음성 인식 시스템의 요구 사항을 해결하기 위하여 은닉 마르코프 모델과 레벨 빌딩 알고리즘을 사용한 인식 방법을 제안한다. 그리고 제안된 인식 시스템의 유용성을 검증하기 위하여 항공기 예약에 사용되는 문장을 정의하고 이 문장을 인식하는 시스템을 설계하였다.

각 단어를 표현하기 위하여 시간에 따라 연속적으로 변하는 신호를 모델링하는데 뛰어난 성능을 가져 음성 인식 분야에 널리 사용되어온 은닉 마르코프 모델<sup>[9-11]</sup>을 사용한다. 인식하고자 하는 모든 문장을 효율적으로 표현하기 위하여 인식 네트워크<sup>[12-15]</sup>를 사용한다. 인식 네트워크는 많은 단어를 표현하는 은닉 마르코프 모델의 집합으로 구성된 문장 정보를 포함한다. 네트워크에서 에지는 단어의 구조적 정보를 표현하고 이는 다른 문장에 의해 구조적으로 공유되고 재사용 될 수 있다. 그리고 초기 상태에서 끝 상태로 도달할 수 있는 경로는 완전한 문장을 표현한다. 이러한 모델에서 인식은 연속 음성 입력패턴이 주어졌을 때 최고의 결합확률(joint probability)을 생성하는 경로를 찾는 문

제로 볼 수 있다.

완전한 문장을 인식하기 위하여 개별적 단어를 표현하는 은닉 마르코프 모델의 연속된 리스트를 평가하는데, 제한된 문법 구조를 가지는 레벨 빌딩(level building) 알고리즘<sup>[7, 8, 13-15]</sup>을 사용한다. 레벨 빌딩 알고리즘은 조합 가능한 문장의 모든 경로를 평가할 경우 발생하는 각 모델의 중복 평가를 피하고, 한 방향으로만 진행하여 인식 네트워크를 보다 빠르게 탐색한다. 각 레벨의 모델들을 탐색할 때는 비터비 알고리즘을 적용하며, 전체 모델에서 인식된 문장의 경로를 역추적할 때는 레벨 탐색시 생성된 최대 확률값을 이용한다. 이러한 레벨 빌딩 알고리즘은 탐색 속도가 단지 네트워크를 이루는 단어를 표현하는 은닉 마르코프 모델의 수에 비례하는 효율적인 네트워크 탐색 알고리즘이다.

본 논문의 구성은 다음과 같다. 제 II 장에서는 단어를 표현하기 위한 은닉 마르코프 모델과 인식 네트워크 탐색에 사용되는 레벨 빌딩 알고리즘에 대한 일반적인 이론을 설명한다. 제 III 장에서는 항공기 예약어 인식을 위한 인식 모델로서 인식 네트워크 및 은닉 마르코프 모델을 제안한다. 제 IV 장에서는 실험을 통하여 제안된 인식 모델의 효용성을 검증하고, 제 V 장에서는 논문의 내용을 요약하고 결론 및 향후 연구 방향에 대해서 논한다.

## II. 은닉 마르코프 모델과 레벨 빌딩 알고리즘

패턴 인식에서 인식 대상은 불확실성(uncertainty), 임의성(randomness)을, 그리고 인식 알고리즘은 불완전성(incompleteness)을 가진다. 이러한 문제를 해결하기 위하여 대표적 통계적 모델인 은닉 마르코프 모델이 많이 사용되고 있다<sup>[9-11]</sup>. 1960년대말 처음 소개된 은닉 마르코프 모델은 Baker와 Jelinek가 처음으로 자동 연속 음성 인식 문제에 적용하였으며<sup>[16]</sup>, 종래의 마르코프 모델로는 해결할 수 없었던 문제들에 성공적으로 적용되고 있다. 패턴 인식을 위한 여러 방법과 비교했을 때 은닉 마르코프 모델은 다음과 같은 장점을 가지며, 이러한 장점들로 인해 음성 인식에 적합함을 보이고 있다.

(1) 지식의 자동 추출: 지식 기반 시스템(Knowledge based system)에서 인식 대상에 대한 지식의

추출은 중요한 단계이다. 은닉 마르코프 모델은 이러한 지식을 데이터들로부터 자동적으로 추출하는 능력을 가지고 있다.

(2) 모델링 능력 : 기존의 구조적 방법과는 달리 패턴들을 학습에 의하여 모델링한다.

(3) 시간적 정보의 모델링 : 은닉 마르코프 모델의 대표적 특징으로서 시간적 정보를 가진 입력 패턴을 효율적으로 표현한다.

(4) 효율적인 학습 및 인식 알고리즘 : 탁월한 모델링 능력을 가지고 있어도 패턴의 학습 시간과 인식 시간이 비현실적이면 실용적인 방법이 아니다. 그러나 은닉 마르코프 모델의 학습을 위한 Baum-Welch 알고리즘, 그리고 인식을 위한 비터비 알고리즘, 레벨 빌딩 알고리즘과 같은 효율적인 알고리즘들이 존재한다.

이 장에서는 한국어 단어 표현을 위해 사용한 은닉 마르코프 모델의 기본 이론을 기술하고, 모델 상에서의 문장 인식 알고리즘으로서, 확장된 비터비 알고리즘 형태인 레벨 빌딩 알고리즘에 대해서 기술한다.

1. 은닉 마르코프 모델

은닉 마르코프 모델은 상태라고 불리는 노드와 이들 간의 전이를 나타내는 에지로 구성된 그래프로 표현할 수 있다. 일반적으로 은닉 마르코프 모델은 아래와 같이 정의할 수 있다.

- (1)  $N$  : 상태의 수
- (2)  $M$  : 관측 심벌의 수
- (3)  $T$  : 관측열의 길이
- (4)  $Q = \{q_t\}$  : 상태 집합,  $q_t \in \{1, 2, \dots, N\}$ ,  $t = 1, 2, \dots, T$
- (5)  $V = \{V_k\}$  : 관측 가능한 심벌들의 집합,  $k = 1, 2, \dots, M$
- (6)  $A = \{a_{ij}\}$  : 상태 전이 확률 분포  
 이는 마르코프 모델에서 상태  $i$ 에서 상태  $j$ 로의 전이 확률을 나타내는 파라미터로,  $a_{ij} = P[q_{t+1} = S_j | q_t = S_i]$   $1 \leq i, j \leq N$  으로 정의할 수 있다.
- (7)  $B = \{b_j(v_k)\}$  : 관측 심벌 확률 분포,  $j = 1, 2, \dots, N$ ,  $k = 1, 2, \dots, M$   
 시간  $t$ 에 상태  $S_j$ 에서  $k$ 번째 심벌  $V_k$ 가 나타나는 확률을 의미한다.
- (8)  $\Pi = \{\pi_i\}$  : 초기 상태 확률 분포,  $i = 1, 2, \dots, N$   
 $\pi_i = P[q_1 = i]$  :  $i$  상태의 초기 확률값.

은닉 마르코프 모델은 이러한 확률 매개변수들을 이용하여 주어진 관측열의 집합을 표현한다. 즉  $O = O_1, O_2, \dots, O_T$  라는 관측열이 주어졌을 때, 외부에서 그 상태 전이 과정을 직접적으로 관측할 수는 없어도 자체의 확률 매개변수를 사용하여 마르코프 과정의 확률 함수로 모델화할 수 있다.

1) 은닉 마르코프 모델 파라미터의 초기 추정치

은닉 마르코프 모델에서 고려해야 할 한 요소로  $A$ ,  $B$  행렬과  $\pi$ 의 초기 확률값 결정을 들 수 있다. 이론적으로 추정식은 우도 함수의 지역 최대값(local maximum)에 상응하는 은닉 마르코프 모델 파라미터의 값을 주어야 한다. 훈련 과정을 통해 추정된 확률값은 지역 최대값이므로 초기값의 변화에 따라 확률값의 변이가 생길 수 있다. 초기값을 설정하는 방법으로는 난수를 사용하는 방법과 동일한 확률값을 주는 방법이 있는데 본 연구에서는 후자의 방법을 사용하였다.

2) 은닉 마르코프 모델의 형태

은닉 마르코프 모델의 형태는 한 상태에서 다른 모든 상태로 전이 확률값을 가지는 에르고딕 모델을 비롯하여 시간이 증가함에 따라 상태의 인덱스도 같이 증가하는 left-to-right 모델(Bakis model) 등 다양한 마르코프 모델이 있다. 이 중 본 논문에서 사용한 그림 1의 left-to-right형 은닉 마르코프 모델은 음성 인식이나 온라인 문자인식에서처럼 시간에 따라 순차적으로 변하는 특성을 지닌 데이터를 모델링하는데 적합한 아래와 같은 특징을 가진다.

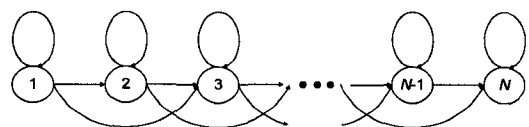


그림 1. Left-to-right형의 은닉 마르코프 모델  
 Fig. 1. A left-to-right HMM.

- (1) 한번 방문한 노드를 다시 방문할 수 없다.

$$a_{ij} = 0, \quad j < i$$

- (2) 상태열은 항상 첫 노드에서 시작하고 마지막 노드에서 종료한다.

$$\pi_i = \begin{cases} 0 & i \neq 1 \\ 1 & i = 1 \end{cases}$$

- (3) 상태 사이의 전이에서 인덱스의 큰 변화는 발생하지 않는다.

$$a_{ij} = 0, \quad j > i + 1$$

이러한 left-to-right 형의 모델을 사용할 경우 고려해야 할 두 가지 사항이 있다<sup>[10]</sup>. 첫째, 하나의 연속적인 긴 관측값은 이 값이 마지막 상태  $q_N$ 에 도달하게 되면 나머지 관측값은 더 이상 이전 상태에 영향을 미치지 못하게 된다. 따라서 이 모델을 훈련시키기 위해서는 첫 상태  $q_1$ 에서 시작되는 관측값들의 집합이 요구되어진다. 둘째, 전향-후향(forward-backward) 확률의 각 개별적 집합에 대한 scale factor가 다르므로 한 패턴에 대한 여러 관측값들을 이용한 추정 과정에서의 scaling에 주의하여야 한다. 이런 문제점을 해결한 다중 관측값을 사용하는 훈련 방법은 다음절에서 설명한다.

### 3) 다중 관측값(Multiple Observation Sequence)

Left-right형 모델의 파라미터에 영향을 미치는 중요한 요소 중의 하나로 훈련을 위한 관측 샘플을 들 수 있다. 본 논문에서는 여러 변이를 가지는 음성 데이터에 독립적인 모델을 구성하기 위해 여러 번 발음된 음성 샘플  $O = [O^{(1)}, O^{(2)}, \dots, O^{(k)}]$ 를 훈련 샘플로 사용한다.  $O^{(k)}$ 는  $k$ 번째 훈련 샘플  $[O_1^{(1)}, O_2^{(2)}, \dots, O_{T_k}^{(k)}]$ 을 나타낸다. 이러한 다중 관측값을 사용하여 모델 파라미터를 추정하는 것은 훈련 샘플들이 서로 독립이라는 가정 하에,

$$P(O|\lambda) = \prod_{k=1}^K P(O^{(k)}|\lambda) \quad (1)$$

$$= \prod_{k=1}^K P_k$$

식 (1)을 최대화 하는 모델  $\lambda$ 의 파라미터를 추정하는 것이다. 파라미터  $A, B$ 를 추정하는 식은 다양한 관측값의 발생 빈도수에 기초를 둔 것이므로 다중 관측값을 위한 추정식도 각 개별적 관측값들의 발생 빈도수를 추가하여 쉽게 확장할 수 있다. 수정된 식은

$$\bar{a}_{ij} = \frac{\sum_{k=1}^K \frac{1}{P_k} \sum_{t=1}^{T_k} a_i^k(j) a_{ij} b_j(O_{t+1}^{(k)}) \beta_{t+1}(j)}{\sum_{k=1}^K \frac{1}{P_k} \sum_{t=1}^{T_k} a_i^k(j) \beta_t^k(j)} \quad (2)$$

와

$$\bar{b}_j(i) = \frac{\sum_{k=1}^K \frac{1}{P_k} \sum_{t=1, s.t. O_t = vt}^{T_k} a_i^k(j) \beta_t^k(j)}{\sum_{k=1}^K \frac{1}{P_k} \sum_{t=1}^{T_k} a_i^k(j) \beta_t^k(j)} \quad (3)$$

이고,  $\pi_1 = 1, \pi_i = 0, i \neq 1$  이다. 이상은 은닉 마르코프

모델에 대한 기본적 설명과 모델의 확률 파라미터의 추정에 대해 기술하였고, 다음은 입력 패턴이 주어졌을 때 이를 사용하여 연속된 은닉 마르코프 모델을 평가하는 레벨 빌딩 알고리즘을 기술한다.

### 2. 레벨 빌딩 알고리즘

일련의 관측값  $O_t, t = 1, 2, \dots, T$ 와 단어를 표현하는 은닉 마르코프 모델의 집합  $\{W_j\}$ , 그리고 단어의 연속으로 표현되어지는 문장이 주어졌을 때 인식의 단계에서는 관측값  $O_t$ 를 사용하여 일련의 적합된 모델  $\{W_{[1]}, W_{[2]}, \dots, W_{[p]}, p = 1, 2, \dots\}$ 을 결정한다. 이는 관측값과 모델 상태의 결합확률을 최고로 하는 정합 문제로 정의될 수 있다.

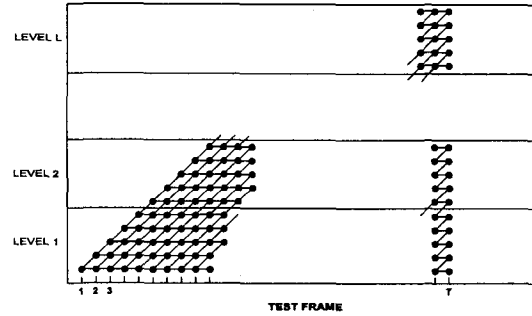


그림 2. 5상태를 가지는 은닉 마르코프 모델에 기초를 둔 level-building의 수행

Fig. 2. Implementation of level building based on five-state HMMs.

그림 2는 전체 모델에서 최고 확률값을 출력하는 일련의 모델을 찾기 위해 적용한 레벨 빌딩 알고리즘을 설명한다. 이 그림에서는 각 개별 모델은 5 상태를 가지며 연속 모델  $W_{[i]}, W_{[j]}, i < j$ 에서는 어떠한 제약 없이 모든 모델이 될 수 있다고 가정한다. 즉, 한 모델 다음에 어떤 다른 모델도 연속으로 배열될 수 있다. 먼저 레벨  $l = 1$ (초기 레벨)의 단계의 프레임 1에서 시작하여 모델  $q(W_q)$ 를 관측값  $O$ 에 정합시킨다. 이러한 초기 레벨의 모델들에 대한 정합값 계산은 하나의 은닉 마르코프 모델에 적용하는 비터비 알고리즘의 초기화(initialization), 순환 계산(recursion), 종료(termination)의 동일한 단계 속에서 수정된 방식에 의해 구해진다. 우선 초기화 과정이

$$\delta_1(1) = [b_1^q(O_1)], \quad (4)$$

$$\delta_1(j) = 0, \quad j = 2, 3, \dots, N \quad (5)$$

의 식들에 의해서 수행된다. 최우항 모델의 특성상 시간  $t=1$  에서 첫 상태에 대한 초기값을 제외한 나머지 상태들에 대한 초기값은 모두 0 이다.

그리고 순환 계산 과정은

$$\delta_i(j) = \max_{1 \leq i \leq N} [\delta_{i-1}(i) * [a_{ij}^q]] * [b_i^q(O_i)], \quad (6)$$

$$2 \leq t \leq T, \quad 1 \leq j \leq N$$

의 식에 의해 수행된다. 여기서  $\delta_i(j)$ 는  $S_i$ 가 시간  $t=i$ 에서의 임의의 상태를 의미할 때, 부분 상태 (partial state)와 관측열의 결합 확률(joint probability)  $\Pr\{O_1, O_2, \dots, O_t \text{ and } S_1, S_2, \dots, S_{i-1}, j \mid A, B\}$  이다. 마지막으로

$$P(l, t, q) = \delta_i(N), \quad 1 \leq t \leq T \quad (7)$$

$$B(l, t, q) = 0 \quad (8)$$

식에 의해 초기 레벨의 각 모델에 대한  $P(l, t, q)$  와  $B(l, t, q)$  를 계산한 후 종료한다.

첫 레벨에서는 모든 모델에 대해 이 알고리즘을 적용하여 계산한 후, 아래의 새로운 배열  $\hat{P}, \hat{B}, \hat{W}$  를 생성한다.

$$\hat{P}(l, t) = \max_q [P(l, t, q)] \quad (9)$$

$$\hat{B}(l, t) = B[l, t, \arg \max_q P(l, t, q)] \quad (10)$$

$$\hat{W}(l, t) = \arg \max_q [P(l, t, q)] \quad (11)$$

위의 식에서  $\hat{P}$  는 각 레벨에서 시간  $t$ 까지의 관측 심벌  $O_1 O_2 \dots O_t$  의 최고 정합 확률을 기록하고,  $\hat{B}$  는 이전 모델에서의 프레임을 가리키는 역추적 포인터 (backpointer)이고,  $\hat{W}$  는 최고 확률을 출력하는 모델을 나타낸다.

레벨 2 이상에서는 초기값을 설정하는 과정에서 레벨 1과 차이점을 가진다. 이들 레벨에서는 이전 단계의 결과를 반영하여 초기값으로 설정해야 하므로 아래와 같이 레벨 1과는 다르게 초기화한다.

$$\delta_1(1) = 0, \quad a_1(1) = 0 \quad (12)$$

$$\delta_i(1) = \max[\hat{P}(l-1, t-1), a_{i1}^q \delta_{i-1}(1)] * [b_i^q(O_i)], \quad 2 \leq t \leq T \quad (13)$$

$$a_i(1) = \begin{cases} t-1 & \text{if } \hat{P}(l-1, t-1) > \delta_{i-1}(1) * a_{i1}^q \\ a_{i-1}(1) & \text{otherwise} \end{cases} \quad (14)$$

식 (12)의  $\delta_i(1)$  의 값은 레벨 2이상의 모델이므로 첫 상태에 첫 관측 심벌  $O_1$  의 결합 확률은 0이 되고, 식 (13)에서  $\delta_i(1)$  는 주어진 모델의 첫 상태에  $t$  번째 관측 심벌  $O_t$  의 결합 확률을 의미하므로, 이 경우  $\delta_i(1)$  는 이전 레벨에서  $t-1$  번째 심벌까지의 결합 확률과 식 (6)에서의  $a_{i1}^q \delta_{i-1}(1)$  의 값 중 큰 값을 가지게 된다. 그리고 이전 모델이 끝나는 전 레벨에서의 프레임의 기록을 위해 식 (14)에서 가장 적절한 초기 역추적 포인터 배열을 생성한다.

순환 계산 동안 역추적 포인터는

$$a_i(j) = a_{i-1}[\arg \max_{1 \leq i \leq N} (\delta_{i-1}(i) * a_{ij}^q)] \quad (15)$$

로 계산되어지며, 레벨의 끝에서 확률과 역추적 포인터 배열은

$$P(l, t, q) = \delta_i(N), \quad 1 \leq t \leq T \quad (16)$$

$$B(l, t, q) = a_i(N), \quad 1 \leq t \leq T \quad (17)$$

로 계산되어진다. 그리고 레벨 1에서와 동일한 방법으로  $\hat{P}, \hat{B}, \hat{W}$  를 결정한 후 다음 레벨로 진행된다. 이 모든 과정은 레벨들 중 최고 레벨  $L$  에 다다르면 끝나게 된다.

레벨  $L$  까지 계산이 끝난 후 각 레벨 단위로 최대 확률을 생성한 모델을 찾는다. 레벨 수가  $L$  일 때, 마지막 시간 확률  $T$  에서  $\hat{P}(L, T)$  을 가지는 모델  $W_{[L]}$  을 결정할 수 있고, 역추적 포인터 배열  $\hat{B}(l, t)$  를 이용하여 전체 관측 심벌에 적합된 모델  $\{W_{[1]}, W_{[2]}, \dots, W_{[L]}\}$  를 구한다. 이와 같은 과정을 수행하는 역추적 알고리즘은 다음과 같다.

**Algorithm (Backtracking)**

```

Begin
    level = L
    back = T
    while (level > 0) do
        begin
            W[level] = Ŵ(level, back)
            back = B̂(level, back)
            level = level - 1
        end
    End
    
```

### III. 제한된 한국어 연속 음성 인식 시스템

인식하고자 하는 단어의 수와 문장의 구조가 제한될 경우의 연속 음성 인식은 서브워드 대신 단어를 인식의 기본단위로 사용하면 단어 내부의 조음결합 문제를 해결할 수 있고 단어별로 모델을 구성하므로 단어사전의 구축이 쉬우므로 보다 실용적인 인식 시스템을 구성할 수 있다<sup>[7]</sup>. 제한된 어휘와 문장 구조를 사용하여 인식 시스템을 구성하는 예로는 열차표 예매, 속담 인식, 항공기 예약, VDS 등이 있으며, 본 논문에서는 단어 단위의 한국어 연속 음성 인식 방법을 적용하기 위하여 간략화된 항공기 예약어 인식 시스템을 사용하였다.

항공기 예약어 인식 시스템이란 전화 혹은 직접 항공사 직원과의 대화를 통해 비행기표를 예매할 경우 사용될 수 있는 문장을 컴퓨터를 사용하여 자동으로 인식하는 시스템이다. 이러한 시스템에서 사용될 수 있는 문장의 종류는 무수히 많지만 실험의 용이성을 위하여 주로 사용되는 문장의 구조를 아래와 같이 제한하였다.

- 3월 14일 서울행 비행기표 있습니까?
  - 5월 7일에 출발하는 서울행 비행기표 있습니까?
  - 8월 30일 부산행 좌석 있습니까?
  - 7월 17일 부산에서 출발하는 제주행 비행기표 있습니까?
  - 다음주 수요일에 출발하는 대구행 좌석 있습니까?
- 이러한 문장을 인식하기 위하여 일반적으로 제한된 단어를 인식하는데 사용되는 은닉 마르코프 모델과 비터비 알고리즘을 사용할 수 있다. 이러한 시스템에서는 각 단어별로 은닉 마르코프 모델을 설계하고 인식 패턴이 주어질 때 이를 각 은닉 마르코프 모델에 대하여 비터비 알고리즘을 사용하여 결합확률을 구하게 되고 이들 중 가장 큰 결합 확률을 생성한 모델을 인식단어로 결정한다.

그러나 이와 같은 방법으로 인식하고자 하는 문장 각각에 대하여 은닉 마르코프 모델로 구성되는 인식 시스템을 설계할 경우는 다음과 같은 문제가 발생한다. 문장 내에서 바뀌는 단어 - 날짜, 도시이름 - 가 제한적이므로 이들이 바뀔 때마다 하나의 큰 은닉 마르코프 모델을 추가한다면 전체 시스템을 위해 필요한 마르코프 모델 수는 매우 많게 된다. 또한 각 마르코프 모델이 긴 문장을 표현해야 하므로 모델의 상대수

도 커지게 된다. 따라서 시스템을 위하여 많은 메모리가 요구되어지고 또한 인식시 많은 상태 수를 가지는 모델에 대하여 비터비 알고리즘을 수행하여야 하므로 많은 연산 시간이 요구되어진다. 이러한 문제점으로 인하여 위의 방법은 사용자와 상호작용이 가능한 인터페이스가 되어야하는 음성 인식 시스템에 부적합하다고 볼 수 있다.

#### 1. 인식 네트워크

본 논문에서는 이러한 문제를 해결하기 위하여 인식 네트워크와 레벨빌딩 알고리즘을 사용한다. 위의 문장은 연속적인 단어열로 이루어진다는 가정 하에 문장에서 단어를 각각 분리하여 에지 위에 표기하고, 이 에지들을 문장이 되도록 연결시킨 상태 전이 다이어그램 형태인 인식 네트워크로 표현될 수 있다. 그림 3은 항공기 예약어 인식 시스템을 위한 인식 네트워크를 보여준다.

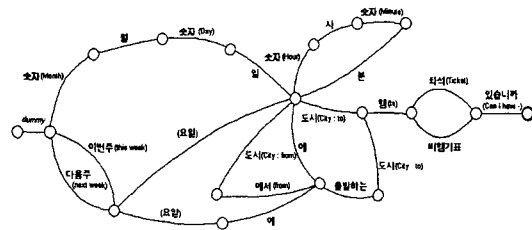


그림 3. 항공기 예약 시스템을 위한 인식 네트워크  
Fig. 3. Recognition network for the airplane reservation system.

이 그림에서 에지 위의 단어는 은닉 마르코프 모델로 표현된다. 다른 에지 위의 동일한 단어는 하나의 마르코프 모델로 표현되고 전체 모델에서 공유되고 재사용되어진다. 따라서 전체 문장을 표현하기 위해서는 구성 가능한 모든 문장에서 사용되는 단어를 추출하여 이를 위해 은닉 마르코프 모델을 구성하면 된다. 그림에서 초기 상태에서 마지막 상태까지의 모든 가능한 전이 경로는 인식하고자 하는 문장에 해당된다. 레벨 빌딩 알고리즘은 입력 음성이 주어졌을 때 이러한 네트워크 구조에서 각각의 마르코프 모델을 평가하고 경로에 따른 확률값을 생성하게 된다. 마지막으로 가장 큰 확률값을 찾고 이 확률을 생성한 경로를 역추적하여 완전한 하나의 문장을 인식하게 된다.

이와 같이 시스템을 구성하게 되면 다음과 같은 면에서 장점을 가지게 된다. 첫째로 음소가 아닌 단어별

로 모델을 구성하므로 단어 내에서 발생하는 조음결합의 문제를 자연스럽게 해결할수 있고, 둘째 새롭게 인식하고자 하는 문장 구조를 추가할 경우와 새 단어를 추가할 경우, 네트워크에서 경로와 에지의 추가로 쉽게 이루어질 수 있다. 즉 단어와 문법 사전의 추가가 용이하다. 그리고 네트워크의 구조적 특성으로 인하여 문장 구조에서 반복되는 단어를 하나의 모델로 표현할 수 있으므로 사전의 크기에 대한 효율성 또한 기대할 수 있다.

2. 전체적인 시스템 구성

본 논문에서 제안된 항공기 예약어 인식 시스템의 전체적인 구성은 다음의 그림 4와 같다. 마이크를 통하여 입력된 아날로그 음성은 A/D converter에 의하여 11Khz, 16bit로 샘플링된 후 끝점 검출기를 통하여 음성인 부분과 음성이 아닌 부분을 구분하게 된다. 음성인 부분에 대하여 10ms씩 이동하면서 20ms단위로 FFT를 수행하고 멜스케일링 기법을 사용하여 연속적인 16차 벡터를 출력하게 된다.

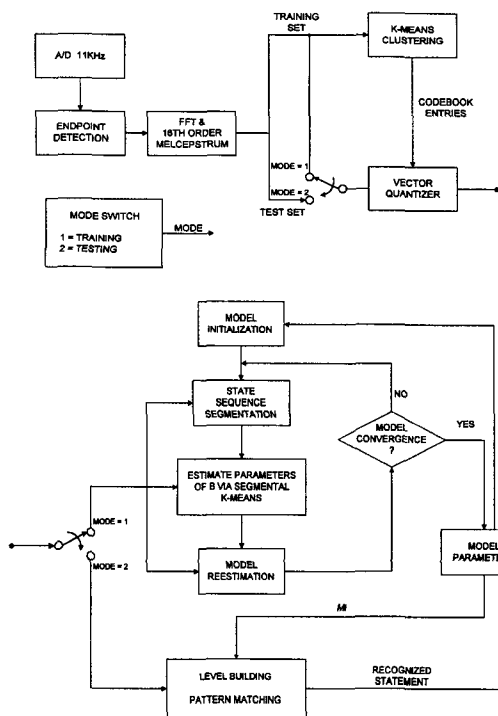


그림 4. 항공기 예약어를 위한 전체적인 시스템 구성  
Fig. 4. System configuration of the airplane reservation.

MODE SWITCH의 값에 따라 시스템은 학습/인

식 모드로 전환된다. 학습 모드(MODE = 1)에서는 연속적인 벡터로 바뀐 음성 신호를 사용하여 코드북을 생성하고 여기에서 생성된 관측심벌을 사용하여 각 단어별로 은닉 마르코프 모델의 확률 파라미터를 추정한다. 인식 모드에서는 훈련에서 생성된 코드북과 입력 신호를 사용하여 관측 심벌을 구한 후 레벨 빌딩 패턴 인식기에서 문장을 인식하게 된다.

1) 끝점 검출과 특징 벡터 추출

마이크로부터 입력된 음성은 11Khz, 16bit로 샘플링되어진 후 끝점 검출기(Endpoint Detection)를 통과하게 된다. 이 끝점 검출기는 입력신호에서 음성인 부분을 가려내는 작업을 수행한다. 이 과정을 거치게 되면 인식을 위하여 처리해야할 신호량이 줄어들게 되고 또한 저장되는 신호량도 줄어들게 된다. 본 논문에서는 STE(short-time energy;  $E(n)$ )와 ZCR(zero-crossing rate;  $Z(n)$ )<sup>[17]</sup> 값을 이용하여 끝점을 검출한다. 이 두 값은 다음과 같이 정의된다.  $s(n)$ 은  $n$  번째 샘플을 의미하고  $N$ 은 윈도우의 크기를 나타낸다.

$$E(n) = \sum_{k=n-N+1}^n s(k)^2 \quad (18)$$

$$Z(n) = \sum_{k=n-N+1}^n |s(k) - s(k-1)| \quad (19)$$

위의 두 값은 녹음할 때 주위의 잡음에 따라 큰 변화를 가지기 때문에 잡음 성분을 제거한 아래의 값을 정의하고 이를 바탕으로 끝점 검출을 수행한다.  $Q_e$ 는 녹음 당시 배경잡음의 log-scale된 STE의 최대치를,  $Q_z$ 는 ZCR의 최대치를 각각 나타낸다.

$$\tilde{E}(n) = 20 \log_{10} E(n) - Q_e \quad (20)$$

$$\tilde{Z}(n) = 20 \log_{10} Z(n) - Q_z \quad (21)$$

그림 5는 입력 파형이 주어졌을 경우 음성의 파형 모양과 끝점이 검출된 영역을 표시하였다.



그림 5. 끝점 검출의 예  
Fig. 5. An example of end-point detection.

끝점이 검출된 음성으로부터 특징 벡터를 추출하는 대표적인 방법으로는 LPC(Linear Predictive Co-

ding)와 FFT(Fast Fourier Transform)가 있다. 실험에서는 20ms에 해당되는 음성 샘플에 대하여 해밍 윈도우를 씌운 후 FFT과정을 수행하게 된다. 해밍 윈도우를 사용하는 이유는 20ms의 샘플에서 양쪽 가장자리 부분의 샘플은 0에서 또는 0으로 급변하는 임펄스 부분에 해당되므로 원 샘플에 존재하지 않았던 고주파 성분이 추가되므로 발생하는 왜곡을 줄이기 위함이다. FFT 수행후 생성된 256개의 주파수 관련 값을 그대로 입력 패턴으로 사용하기에는 데이터 양이 너무 크므로 멜스케일링 기법을 이용하여 고주파 영역에 비하여 저주파 영역이 더 세밀하게 분할되도록 하여 16개의 차수를 가지는 벡터로 데이터를 축소시킨다.

## 2) 벡터 양자화(코드북 생성)

이산 은닉 마르코프 모델을 사용할 경우 모델의 입력 패턴은 유한개의 클러스터 중 하나의 값을 가져야 한다. 이전 단계에서 생성된 16차 벡터는 이산치가 아니어서 이산 은닉 마르코프 모델에 직접 사용할 수 없으므로 이를 유한개의 클러스터로 나누는 과정이 꼭 필요하게 된다. 클러스터링 방법으로는 K-Means 알고리즘<sup>[17]</sup>을 사용하였다. 이 알고리즘은 클러스터의 개수  $K$ 가 주어질 경우, 임의로 전체 벡터 중에서  $K$ 개의 중심을 설정한다. 그 후 모든 벡터 각각에 대하여 각 중심과의 거리를 계산하여 어떠한 클러스터에 속하는지를 결정하게 된다. 모든 벡터를  $K$ 개의 클러스터에 나눈 후 각 클러스터별로 새로운 중심을 구하게 된다. 이러한 과정을 여러 번 반복 수행하면서 각 클러스터간의 중심 사이의 유클리디안 거리가 점점 멀어지게끔 분할하게 된다.

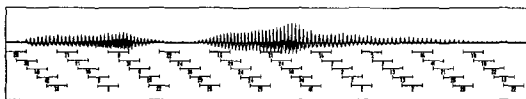


그림 6. 단어 '대구'의 관측심벌  
Fig. 6. Observation symbol of a word '대구'.

그림 6은 '대구'라는 단어를 발음하였을 경우, 48개의 클러스터를 가지는 코드북을 사용하여 파형으로부터 얻어지는 코드열을 나열한 것이다. 그림에서 나누어 있는 20ms의 작은 구간 단위로 특징 벡터를 추출하고 코드를 구한다. 그리고 10ms 간격으로 윈도우를 중첩시키며 이동하여 연속적인 코드를 얻는다. 각 코드 값은 20ms의 음성이 48개의 서로 다른 특징들 중 가장 가까운 특징 벡터를 가리키는 값이라고 볼 수 있

다. 이렇게 생성된 코드를 은닉 마르코프 모델의 관측 심벌로 사용하여 모델을 훈련, 평가하게 된다.

## 3) 단어를 위한 은닉 마르코프 모델의 생성 및 훈련

인식 네트워크에 나타난 단어를 표현하기 위하여 음성 인식 분야에 널리 사용되는 그림 1의 left-right형의 은닉 마르코프 모델을 사용하였다. 이 모형은 시간에 따른 상태의 변화가 선형적인 데이터를 모델링하는데 적합한 특성을 보이므로 음성과 같은 데이터를 잘 표현할 수 있다. 그리고 개별적 단어 각각을 표현하기 위하여 단어별로 하나의 은닉 마르코프 모델을 설계하였다. 각 모델은 4~10 상태를 가지며 각 상태에서의 관측 심벌은 코드북에서 선택된 값 중 하나로 결정된다.

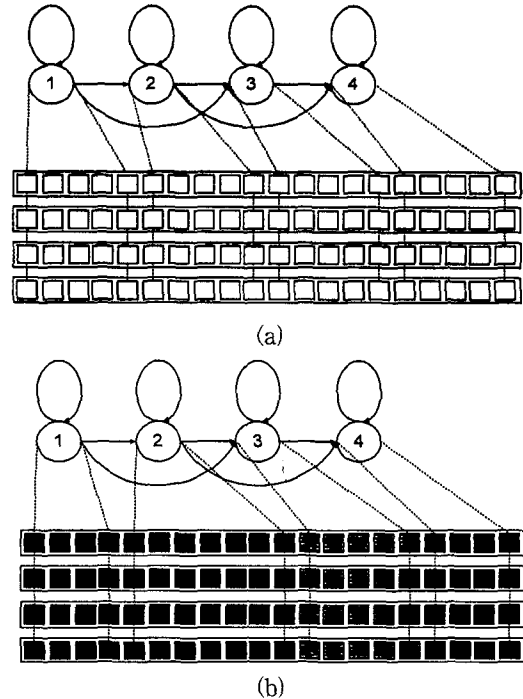


그림 7. 입력심벌을 은닉 마르코프 모델의 상태로 분할

Fig. 7. Input symbol mapping to the state of HMM.

그림 4의 시스템 구성도에서 MODE SWITCH가 1일 경우 각 단어 표현을 위한 은닉 마르코프 모델이 학습된다. 학습과정은 다음의 단계로 수행되어진다. 먼저 모델 초기화 과정에서 확률 파라미터  $A$ 와  $B$ 는 동일한 확률값이 주어진다. 그 후 동일 단어에 대한  $k$ 개의 훈련 샘플이 주어질 경우, 일반적으로 은닉 마르코프 모델의 상태수 보다 관측 심벌열의 길이가 크므로



이를 그대로 사용하여 left-right형의 모델을 훈련시키게 되면 지역 최대값(Local Maximum)에 빠지게 되어 관측열의 뒷부분을 정확히 모델링할 수가 없다<sup>[10]</sup>.

이러한 문제를 해결하기 위하여 상태열의 분할(State Sequence segmentation)과정과 K-Means 분할 알고리즘을 사용하여 확률 파라미터  $B$ 를 추정한 후 은닉마르코프 모델을 재추정할 수 있다. 그림 7의 (a)와 같이 시간에 따라 입력 심벌을 마르코프 모델의 상태수에 맞게 동일하게 분할한다. 그리고 난 뒤 네 개의 분할의 중심을 구하게 되고 K-Means 알고리즘을 사용하여 이들 중심을 재추정하고 네 개의 클러스터로 분할한다. 그림 7의 (b)는 상태에 따라 분할된 심벌을 나타내고 있다. 분할된 각 클러스터의 값을 평균 내어 은닉 마르코프 모델의 확률 파라미터  $B$ 의 초기값으로 할당한다. 이와 같은 과정을 거치게 되면 지역 최대값으로 잘못 추정되는 경우를 막을 수 있다.

이렇게 확률 파라미터  $A$ 와  $B$ , 그리고  $k$ 개의 관측 심벌의 열이 주어지면 2장의 다중 관측값을 이용한 파라미터의 재추정 식을 사용하여 마르코프 모델을 훈련시킨다. 모델의 훈련을 마치고 나면 추정된 확률 파라미터  $A$ 를 다시 그 모델의 초기확률로 설정하고 파라미터  $B$ 를 위의 방법으로 추정한 후 모델을 다시 훈련시키게 된다. 훈련이 끝나게 되면 이전에 추정한 은닉 마르코프 모델과 새로 추정된 은닉 마르코프 모델을 비교하여 그 값의 차이가 문턱값(Threshold)이하이면 추정과정을 멈추고 그렇지 않으면 위의 과정을 반복하여 추정하게 된다.

4) 레벨 빌딩을 이용한 인식 네트워크의 탐색

문장에서 분리된 작은 단위인 단어를 독립된 은닉 마르코프 모델로 표현함으로써 문장 전체를 학습하는 경우에 비해 충분한 학습 샘플을 얻을 수 있으므로 보다 정확한 은닉 마르코프 모델의 확률 매개변수를 추정할 수 있다. 그러나 전체 문장을 인식해야할 경우 분리된 모델 사이의 전이확률이 존재하지 않으므로 기존의 비터비 알고리즘으로는 모델을 평가할 수 없다.

본 논문에서는 분리된 모델별로 입력된 관측 심벌의 열을 정합시킬 수 있는 레벨 빌딩 알고리즘을 제한된 네트워크 구조에서 사용될 수 있도록 수정 적용하였다. 이 알고리즘은 각 모델에서 시간의 변화에 따른 관측 심벌의 정합 확률과 역추적 경로를 저장하여, 전체 모델에서 최적의 확률을 생성한 경로를 유지한다. 그리고 전체 네트워크를 구성하는 개별적 모델의 수에 비

례하지 않고 각 레벨에서의 모델 수에 비례하는 빠른 탐색 알고리즘이다.

제 2.2 절의 레벨 빌딩 알고리즘은 한 모델 다음에 다른 어떤 모델도 연속할 수 있다는 가정 아래 아무런 제약이 없는 알고리즘이다. 그러나 그림 3의 인식 네트워크는 경로에서 표시되는 제한된 문장만을 인식하므로 2.2절의 레벨 빌딩 알고리즘 대신 제약 조건 - 레벨 빌딩 알고리즘에서  $l$ 번째 레벨에서의 선택 가능한 후보는  $(l + 1)$ 번째 레벨에서의 연속적으로 후보가 될 필요는 없다 - 을 가지는 레벨 빌딩 알고리즘을 필요로 한다. 즉 그림 3에서 '월'을 표현하는 은닉 마르코프 모델 다음에는 숫자를 표현하는 모델만이 올 수 있으므로 네트워크의 탐색 공간을 줄일 수가 있다.

각 레벨은 주어진 시간  $t$ 에서 최고 확률을 생성하는 단어 모델을 출력하고, 마지막 단계에서 역추적 과정을 통해 적절한 단어 모델의 열을 구한다. 수정된 레벨 빌딩 알고리즘은 매 레벨마다 인식 네트워크에 존재하는 모든 모델에 비터비 알고리즘을 적용하지 않고 가능한 경로 내에 있는 모델에만 적용한다. 다음은 간단한 인식 네트워크를 설계하고 이에 제한된 레벨빌딩 알고리즘을 적용한 결과를 설명한다. 그림 8은 '○월 ○일 출발하는 ○○'을 표현하는 간단한 인식 네트워크이고, 전체 레벨의 수는 4이고 네트워크를 구성하는 은닉 마르코프 모델의 수는 12개이다. 전체 표현 가능한 문장의 수는  $48(4 \times 4 \times 3)$ 개에 해당된다.

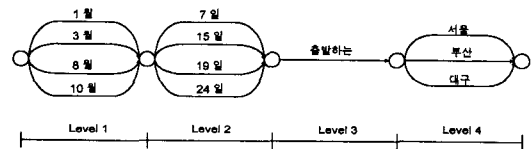


그림 8. '○월 ○일 출발하는 ○○'을 표현하는 인식 네트워크  
Fig. 8. Recognition network for '○월 ○일 출발하는 ○○'.

레벨 1에서는 월(1월, 3월, 8월, 10월)을 표현하는 각 모델  $W_i$ 에 대해서 위의 수정된 비터비 알고리즘을 적용한 후 이들 은닉 마르코프 모델 중에 각 프레임(관측값,  $O = O_1, O_2, \dots, O_T$ )별 최고 확률값을  $P(1, t)$ 에 저장하고, 최고 확률을 출력하는 모델의 인덱스를  $\hat{p}(1, t)$ 에 저장한다. 첫 레벨이므로 역추적 포인터는 어떤 모델도 가리키지 않는다. 이들은 각 시간  $t$ 에서 최적합 모델을 결정하여 다음 레벨의 비터비 계

산의 초기값으로 사용된다. 식 (13), (14)에 의해 이전 레벨에서 최고 확률값과 현재 레벨에서 모델의 첫 상태에 머무를 확률값을 비교하여 이전 레벨에서의 연결 위치와 프레임을 기록하는 초기화 과정을 수행한다. 초기화한 후 비터비 알고리즘의 순환 과정을 수행하면서 각 모델의 각 상태에서 가장 큰 확률값을 프레임에 기록하여 역추적 포인터로 사용한다. 모든 레벨에 대한 비터비 계산이 끝난 후 적합 모델을 결정하는 역추적 과정을 수행한다.

으며 총 녹음된 1400개의 문장을 실험에 사용하였다. 그림 9의 (a)는 입력 음성으로부터 특징 벡터를 추출하여 벡터 양자화 과정을 거쳐 코드북을 생성하는 프로그램이고, (b)는 문장이 주어졌을 때 레벨빌딩 알고리즘을 사용하여 인식하는 프로그램이다.

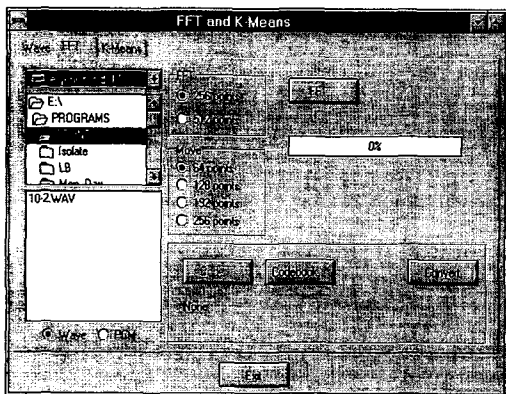
본 논문의 항공기 예약어 인식 시스템에서 인식하고자 하는 문장을 구성하는 단어는 표 1과 같다.

표 1. 인식 시스템에 사용된 단어  
Table 1. Words in the recognition system.

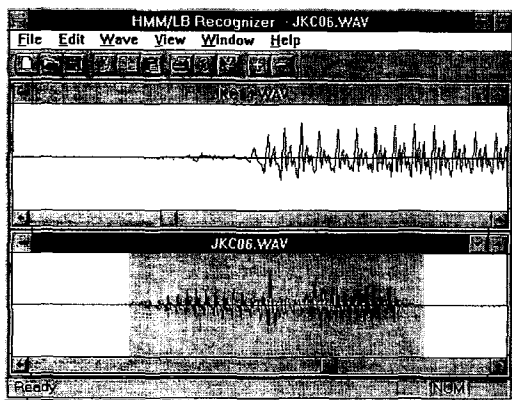
항공기 예약어 인식 시스템을 구성하는 단어		
월	이	육십
일	삼	서울
시	사	부산
분	오	대구
이번주	육	제주
다음주	칠	인천
에, 에서	팔	월요일
출발하는	구	화요일
행	십	수요일
좌석	이십	목요일
비행기표	삼십	금요일
있습니까	사십	토요일
일	오십	일요일

IV. 실험 및 결과

실험에 사용한 데이터는 20명의 여성화자로부터 일 반적인 잡음이 존재하는 실험실에서 녹음하였다.



(a)



(b)

그림 9. 인식 프로그램 레이아웃  
Fig. 9. Layout of the recognition program.

인식 시스템은 Pentium-90 PC 상의 Windows NT 3.51 환경 하에 C++ 언어를 사용하여 구현하였

이들 단어를 은닉 마르코프 모델을 사용하여 표현할 때 중요한 것은 마르코프 모델의 상태수와 관측 심볼의 종류가 되는 코드북의 클러스터의 수이다. 일반적으로 화자 독립 시스템을 구현할 경우는 128개 또는 256개를 사용한다<sup>[19]</sup>. 그러나 본 연구에서는 인식 네트워크의 크기 및 실험 단어의 수가 많지 않으므로 적은 수의 코드북을 사용하는 것이 보다 효율적이었다. 표 2는 숫자(1에서 10)에 대한 코드북의 크기와 은닉 마르코프 모델의 상태 수에 대한 인식률을 보여준다. 은닉 마르코프 모델의 상태수가 6이고 코드북의 크기가 48개일 경우 가장 높은 단어 인식률 98.0%를 보였다. 모델의 상태수나 코드북의 크기를 결정하는 수학적 접근법이 없으므로 이러한 결정은 실험을 통해서만 가능하다<sup>[10]</sup>.

전체 1400개의 문장 중 15명의 화자로부터 추출한 1050개의 문장을 마르코프 모델의 훈련에 사용하였고 나머지 5명으로부터 추출한 350개의 문장을 테스트에 사용하였다. 훈련에 사용한 데이터에 대한 인식률은 94.4%였고, 테스트에 사용한 데이터에 대한 인식률은

90.8%였다. 여기에서 인식률은 실험에 사용된 문장 내에 사용된 단어 수와 인식한 단어 수의 백분율이다. 표 3은 오인식된 단어의 종류와 오인식률을 보여준다.

표 2. 코드북의 크기와 모델의 상태수에 따른 인식률  
Table 2. Recognition rate by codebook size and the number of state.

크기 상태수	32	48	64	128	256
4	81.2	91.3	90.2	91.7	88.0
5	82.5	95.3	92.0	92.4	89.0
6	82.0	98.0	93.2	92.3	88.4
7	81.0	96.2	91.5	90.0	87.7
8	80.0	94.4	89.4	89.8	85.7

표 3. 단어별 오인식률  
Table 3. Ill-recognition rate based on word.

단어	학습용 데이터		실험용 데이터	
	#	%	#	%
숫자	334 / 2280	14.6	112 / 516	21.7
도시	136 / 1256	10.8	45 / 316	14.2
요일	99 / 788	12.6	31 / 184	16.8
나머지	121 / 8004	1.5	82 / 1924	4.3
전체	690 / 12328	5.6	270 / 2940	9.2

표에서 보는바와 같이 오인식은 주로 많은 가지수를 가진 단어인 숫자, 도시 이름, 요일명 등에서 발생한다. 그 이유는 인식할 문장의 문법적 구조가 미리 인식네트워크에서 규정되기 때문에 번이가 많이 발생할 수 있는 숫자, 요일에서 오인식이 주로 발생한다. 특히 한국어 숫자음은 숫자음 사이에 모호함이 존재하므로 많은 오인식을 일으킨다. 그 예로 17 /siptsil/ 과 7 /tsil/을 발음할 경우, 이 두 단어의 첫 번째 발음이 매우 유사하므로 구분하기에 어려움을 지닌다. 음성학적으로 /s/ 발음(마찰음)과 /ts/ 발음(파열음) 사이의 구분은 모호하다.

V. 결론

한국어 음성 인식에 대한 연구는 주로 음절 또는 음소와 같은 서브워드를 인식의 기본 단위로 사용하여 수행되어져 왔다. 그러나 예약 시스템 또는 VDS와 같은 제한된 언어의 사용을 요구하는 시스템일 경우,

서브워드를 사용하는 것보다는 조음결합 문제를 줄일 수 있는 단어를 인식의 기본단위로 사용하는 것이 실용적이다.

본 논문에서는 단어 단위의 한국어 연속 음성 인식 방법을 제안하였다. 인식 대상은 항공기 예약어라는 축소된 범위의 연속 음성이다. 항공기 예약에 필요한 각 단어들은 은닉 마르코프 모델을 이용하여 표현하였고, 이러한 단어들로 구성되는 유효한 문장들은 인식 네트워크를 이용하여 모델링 하였다. 모델 상에서 입력 음성을 인식하기 위한 최적의 경로는 레벨 빌딩 알고리즘으로 탐색하였다. 실험에서 높은 인식률과 빠른 인식 속도를 보였으며, 비교적 적은 저장공간으로 전체 문장을 표현하였다. 또한 새로운 문장 구조의 도입 시에 쉽게 확장할 수 있다는 장점을 가지고 있다. 이는 실용적인 연속 음성인식 시스템의 가능성을 제시한다.

제안한 인식 시스템을 개선하기 위한 향후 연구 과제는: (1) 오인식의 주류를 이루고 있는 숫자음을 인식하기 위한 보다 강력한 특징 추출; (2) 음성을 연속적으로 표현하기 위한, 이산적 관측 심별이 아닌 연속 분포의 도입; (3) 그리고 어휘 수를 확장하기 위한 보다 효율적인 데이터베이스의 구축에 관한 연구이다.

참고 논문

- [1] 구명완, "음성인식의 현황," 대한전자공학회지, 제20권, 제5호, pp. 41-50, 1993
- [2] 김경희, 이근배, 이종혁, "한국어 음성 언어 처리를 위한 음소 단위 인식과 형태소 분석의 결합," 한국정보과학회 논문지, 제22권, 제10호, pp. 1488-1498, 1995
- [3] 이영호, 정홍, "음절을 기반으로한 한국어 음성 인식," 대한전자공학회 논문지, 제31권, 제1호, pp. 11-22, 1994
- [4] 김동국, 정차균, 정홍, "한국어 음소 인식을 위한 시간 지연 신경망," 한국정보과학회 논문지, 제18권, 제3호, pp. 300-312, 1991
- [5] 박규봉, 이근배, 이종혁, "음소단위 TDNN에 기반한 한국어 연속 음성 인식을 위한 데이터 자동분할," 제7회 한글 및 한국어 정보처리학회, 제16권, pp. 78-82, 1995
- [6] H. J. Yu and Y. H. Oh, "A Fuzzy Expert System for Continuous Speech Recognition using Non-Uniform Unit,"

- Journal of KISS*, vol. 22, no. 1, pp. 128-135, January 1995.
- [ 7 ] L. R. Rabiner and S. E. Levinson, "A Speaker-Independent, Syntax-Directed, Connected Word Recognition System Based on Hidden Markov Models and Level building," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 30, no. 4, pp. 561-573, June 1982.
- [ 8 ] C. S. Myers and L. R. Rabiner, "A Level Building Dynamic time Warping Algorithm for Connected Word Recognition," *IEEE Trans. Acoust. Speech, signal Processing*, vol. 29, no. 2, pp. 284-297, April, 1981.
- [ 9 ] L. R. Bahl, F. Jelinek, and R. L. Mercer, "Maximum Likelihood Approach to continuous Speech Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 5, no. 2, pp. 179-190, 1983.
- [ 10 ] L. R. Rabiner, "Tutorial on Hidden Markov Model and Selected Application in Speech Recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257-285, 1989.
- [ 11 ] L. R. Rabiner, S. E. Levinson, and M. M. Sondhi, "An Introduction to the Application of the Theory of probabilistic Function of a markov Process to Automatic Speech Recognition," *The Bell System Technical Journal*, vol. 62, no. 4, pp. 1035-1074m, 1983.
- [ 12 ] 신봉기, 김진형, "은닉 마르코프 모델 네트워크에 의한 온라인 홀림 필기 한글 인식," 한국정보과학회 논문지, 제21권, 제9호, pp. 1737-1745, 1994
- [ 13 ] H. J. Kim, K. H. Kim, and S. K. Kim, "On-line Recognition of Handwritten Chinese Character Based on Hidden Markov Models and level building," *Pattern Recognition*, vol. 30, no. 9, pp. 1489-1500, 1997.
- [ 14 ] K. H. Kim, S. K. Kim, J. G. Lee, and H. J. Kim, "On-line Recognition of Cursive Korean Characters Based on Hidden Markov Model and level Building," *KITE Journal of Electronics Engineering*, vol. 7, no. 2, pp. 60-70, June 1996.
- [ 15 ] K. H. Kim, S. K. Kim, and H. J. Kim, "Recognition of continuous Spoken Korean Language using HMM and Level Building," *Proceedings of International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC'96)/Seoul*, vol. 2, pp. 1071-1074, 1996.
- [ 16 ] J. K. Baker, "The DRAGON System - an Overview," *IEEE Trans. Acoustics, Speech, and Signal processing*, vol. 23, no. 1, pp. 24-29, 1975.
- [ 17 ] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*.
- [ 18 ] J. S. Pan, F. R. McInnes, and M. A. Jack, "Fast Clustering Algorithms for Vector Quantization," *Pattern Recognition*, vol. 29, no. 3, pp. 511-518, 1996.
- [ 19 ] A. M. Peinado, J. C. Segura, A. J. Rubio, P. Garcia, and L. Perez, "Discriminative Codebook Using Multiple Vector Quantization in HMM-Based Speech Recognizers," *IEEE Trans. Speech, and Audio processing*, vol. 4, no. 2, pp. 89-95, 1996.
- [ 20 ] K. F. Lee, "Context-Dependent Phonetic Hidden Markov models for Speaker-Independent Continuous Speech Recognition," *IEEE Trans. Acoustics, Speech, and Signal processing*, vol. 38, no. 4, pp. 599-609, 1990.
- [ 21 ] K. F. Lee, H. W. Hon, and R. Reddy, "An Overview of SPHINX Speech Recognition System," *IEEE Trans. Acoustics, Speech, and Signal processing*, vol. 33, no. 1, pp. 35-45, 1990.

저자 소개



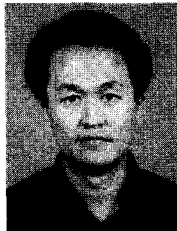
金庚炫(正會員)  
1973년 5월 30일생. 1995년 2월 경북대학교 통계학과 졸업(이학사). 1997년 2월 경북대학교 대학원 컴퓨터공학과 졸업(공학석사). 1997년 3월~현재 현대중공업 마북리 연구소 근무. 주관심분야는 패턴인식, 데이터

베이스 등임



金相均(正會員)  
1965년 8월 16일생. 1991년 8월 경북대학교 통계학과 졸업(이학사). 1994년 2월 경북대학교 대학원 컴퓨터공학과 졸업(공학석사). 1996년 8월 경북대학교 대학원 컴퓨터공학과 졸업(공학박사). 1996년 9월~현재

인제대학교 정보컴퓨터학부 전임강사. 주관심분야는 패턴인식, 영상처리 등임



金恒俊(正會員)  
1954년 12월 30일생. 1977년 2월 서울대학교 전기공학과 졸업(학사), 1979년 KAIST 전기 및 전자공학과 졸업(공학석사). 1997년 시즈오카대학교 전자과학연구과 졸업(공학박사). 1980년~현재 경북대학교 컴퓨

터공학과 교수. 주관심분야는 인공지능, 영상처리, 패턴인식, 병렬처리, 컴퓨터구조 등임