연속 분포 HMM에 의한 실시간 Word Spotting에 관한 연구

서상원*, 박전규*, 이종현*, 김도석**, 한문성*
*한국과학기술연구원 시스템공학연구소, **한국과학기술원 전기및전자과

A Study on the Real-time Word Spotting by Continuous density HMM

Sangweon Suh*, Jeon-Gue Park*, Chong-Hyun Lee*, Don-Suk Kim**, Mun-Sung Han*
*Systems Engineering research Institute/KIST, **Dept. of Electrical Eng./KAIST

요 약

본 논문에서는 연속 분포 HMM을 사용한 실시간 로보트 암 제어 시스템에 대해 기술하고 있다. 본 시스템은 자연스러운 문장의 로보트 암 제어 명령 발성을 받아 핵심단어 인식의 framework을 통한 명령 인식 및 로보트 제어를 구현하고 있다. 로보트 몸체의 부분, 방향, 각도, 동작명령들에 대해 각기 우향(left-to-right) HMM, 이외의 비 핵심어들에 대해서는 이들을 한데 모아 ergodic형 상태천이를 모델링하는 garbage HMM을 형성했는데, 조사, 감탄사 등을 따로 모은 garbage 모델과, silence 및 배경 잡음에 대한 garbage 모델을 형성, 학습 및 인식에 포함시켜 연결단어 인식을 수행함으로써 핵심단어 인식의 효과를 얻었다. 이때 핵심단어들의 사용에 있어 간단한 문법적 제약을 가정하였다. 남성화자 35명을 대상으로 30개 문형에 대해 데이타 수집용 개념적 문장을 구성하여 음성 데이타를 수집하였다. 학습 화자에 대한 제어 명령 인식률은 95%이상을 나타내고 있으며, 비 학습 화자에 대한 인식율은 90%이상이다. 또한 학습된 단어외의 비 핵심단어들의 사용에 대해서도 긍정적인 인식 성능을 보였다.

1 서 론

본 시스템은 전형적인 음성 명령어 인식시스템의 하나인 로보트 제어를 위한 음성 인터페이스를 공중전화망을 대상으로 구현한 시스템으로서, 전화회선을 통해 격리된 지역에 존재하는 전자적 기기나 설비를 원격조종할 수 있도록 하는 시스템이다.

현재 전형적인 ARS(Audio Response System)의 일종으로서, 전화기의 DTMF 신호를 통해 가전기기나 산업기기를 제어하는 시스템이 이미 상용화되어 사용중에 있는데, 이러한 시스템의 주요 단점은 필요할 때마다 사용자가 제어 코드를 기억해서 해당 제어코드와 관련된 번호를 버튼을 통해 입력시키는 방식으로서, 자연성이 없으며 입력속도가 느린 단점이 있다. 또한 전화를 받는 ARS 서버시스템에서 제공하는 획일적인 시나리오에 의해 운영되기때문에 사용자는 대화과정에 적극적으로 참여할 수가 없게 되며 따라서 거부감이 생기게 된다.

음성은 그 자체의 자연성으로 인해 차세대 사용자 인터페이스로서 이미 충분한 가능성을 입증받고 있는데, 미국이나 일본은 이미 전화망을 대상으로 하는 응용시스템을 개발하고 부분적인 실험 및 서비스를 개시하고 있으며, NTT의 ANSER[3], AT&T의 전화번호인식 시스템[9] 등이 대표적인 시스템이다. 또한 발화된 연속음성중의 keyword를 추출함에 의해 음성인식의 응용을 현실화하는 기술이 있는데 TOSBURG[8]와 같은 시스템이 있다.

이러한 배경을 기반으로 본 논문에서는 전화망을 대상으로 연속 HMM에 의한 keyword spotting기법을 구현하고 추출된 keyword를 해당 로보트 제어신호로 변환하여 로보트 구동을 실현하는 시스템에 대해 기술하고 있다.

2 시스템의 개요

본 시스템은 그림 1에서와 같은 구조로 되어 있는데, PC와 Digital Signal Processing(DSP) 보드 및 PC와 serial port를 통해 전달되는 제어신호에 의해 작동하는 로보트 암으로 구성되어 있다. 사용자는 로보트 시스템과 대화를 하듯이 로보트를 제어할 수 있는 잇점이 있으며, keyword spotting 기법에 의해 발화문장으로부터 로보트 제어에 필요한 어휘를 추출하고 있다. keyword spotting 기법에 의해 추출된 제어에 관계된 명령어는 PC에 설치된 관리자 프로그램에 의해 제어신호가 발생, 이를 실제 로보트 시스템에 전달해서 로보트가 구동되는 것을 실연하고 있다. 이를 위해 사용자와 전화벨을 감지하고 통화를 수립하는 한편 PC와 신호처리보드간의 양방향 통신을 수립하여 필요한 자원들을 다수의 신호처리보드가 공유하도록 하고 있다.

소프트웨어는 크게 두개의 부분으로 구분되는데 PC 관리자 프로그램과, DSP 관리자 프로그램으로 구성된다. PC 관리자 프로그램은 다시 DSP 보드와의 통신 수립과 네개의 DSP 보드의 다중처리를 지원하는 주 프로그램, 로보트의 동작을 그대로 화면상에 제시하는 그래픽 사용자 인터페이스 부분, 사용자가 원하는 로보트에 원하는 제어신호를 발생시킬 수 있도록 하는 사용자 선택형 로보트 기기 제어부분, 로보트를 실제 구동하도록 제어 신호를 발생시키고 serial port를 통해 전달하도록 하는 제어신호 발생 및 전송부 등 네 개의 하부 구조로 구성되어 있다.

DSP 보드상에서 수행되는 프로그램은 다섯 개의 부분으로 구성되어 있는데 PC 관리자와의 통신을 전담하고 전화벨 감지후 사용자와의 회선을 수립하는 주 프로그램, 통화가 일단 수립되면 상시 사용자의 음성이나 명령을 감지하도록 하는 사용자에 의해 호출되는 로보트의 이름 인식부, 음성을 녹음하고 재생하는 부분, 녹음된 음성으로부터 신호처리를 통해 음성 특성을 추출하는 부분, 마지막으로 본 시스템의 핵심 부분인 연속 HMM에 의한 음성인

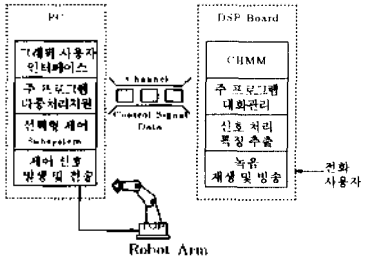

그림 1: 로보트 제어를 위한 음성 인식 시스템의 구조

식 부분이 그것이다.

3 연속 분포 HMM을 이용한 단어별 학습

그림 2는 본 시스템에서 사용한 문법을 Finite State Network(FSN)으로 도해하고 있는데, 원안와 문자는 인식 단어 즉 핵심 단어를 의미하며 원밖의 문자는 각각 S#는 silence, G#은 garbage 모델을 의미한다.

학습은 각 핵심단어별 데이타, garbage 단어 집합의 데이타, 그리고 silence 및 background 잡음의 데이타에 대하여 단어별 HMM 학습을 수행하였다. 학습 알고리즘은 표준적인 EM 알고리즘을 사용하였다. HMM의 forward-backward 절차에의한 학습 과정은 잘 알려져있다[5]. HMM의 상태 확률 모델은 continous mixture density를 사용하였다.

$$b_j(O) = \sum_{m=1}^{M} c_{jm} N(O \mid \mu_{jm}, \Sigma_{jm})$$

$N(O \mid \mu_{jm}, \Sigma_{jm})$은 정규분포의 확률밀도 함수를 나타낸다($j = 1, \cdots, N$ $m = 1, \cdots, M$).

$$(2\pi)^{-\frac{p}{2}} \mid \Sigma_{jm} \mid^{-\frac{1}{2}} \exp\{-\frac{1}{2}(O - \mu_{jm})' \Sigma_{jm}^{-1} (O - \mu_{jm})\}$$

M은 mixture의 갯수, N은 단어 HMM의 상태의 갯수이며 p는 특징 벡터의 차원을 나타낸다.

상태 천이 확률에 있어 silence 및 garbage 모델의 경우 상태간의 full communication을 가정하였으며 핵심단어 모델의 경우 우향 HMM 구조를 채택하였다. 그림 3은 5 상태 silence HMM과 핵심단어인 '어께'에 대한 6 상태 HMM의 학습 결과 추정된 상태 천이 확률들을 보여준다. silence HMM의 상태 천이 구조 추정은, 온닉 마르코프 프로세스의 특성에 있어 유한 개의 모든 상태들간의 communication이 있고 주기성이 없으므로 stability를 가지며 대응하는 stationary 초기 상태 분포를 사용할 경우 ergodic 프로세스가 되는 모델을 형성한 것이다. 학습으로부터 얻어진 단어별 모수 집합을 위에서 설명한 FSN에 따라 loading 하여 연결 단어 인식을 수행한다.

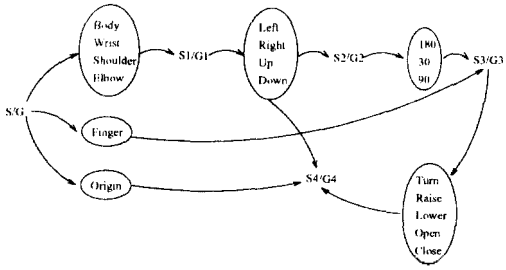

그림 2: 로보트 암 제어를 위한 grammar network의 구성

4 탐색 알고리즘의 구현

인식 탐색 알고리즘은 단어 인식을 위한 Viterbi 탐색의 단어 level을 포함한 연장으로서 프레임 동기 one-pass 알고리즘[6]을 구현하였다. 연결된 단어들로서의 연속 음성 인식의 방법론은 보통 인식을 위해 사용되는 모든 지식들(예, 단어 표현, 언어 모델)을 확률적인 또는 결정적인 network로 나타내고, 음성을 나타내는 subword 또는 단어 모델과 기본 network와 결합하고, 전체 network을 Dynamic Programming을 이용하여 효과적이고 정확하게 탐색할 수 있다는 생각에 기반을 둔다.

이 문제를 풀기위한 여러 알고리즘들이 개발 발전되어 왔는데, Stark 알고리즘[1], Two Level DP Matching[7], Level-Building 알고리즘[2], one-stage DP 알고리즘[4], 프레임 동기 DP 탐색 알고리즘[6] 등이 알려져 있다. 이들은 모두 문법적 제약하에서 최적의 단어열을 구하는 기능을 가지고있다. 이들 탐색 알고리즘들의 주된 차이점은 프레임 동기나 단어 동기를 적용하는 알고리즘 구현상의 특색들이다.

단어인식을 위한 Viterbi 탐색으로부터 한 레벨 더 확장하여 매 시간 t에서 단어와 단어간의 천이를 함께 고려하는 탐색을 하는 것이 기본적인 생각이다. 이에 따른 Viterbi 탐색은 가능한 duration과 천이를 가진 모든 단어열 중에 가장 높은 우도 점수를 갖는 또한, 그 열와 각 단어 안에서는 각기의 HMM 단어 모델에 따라 optimal 상태열을 갖는 열을 탐색해 내는 문제로서, 2-레벨 Dynamic Programming의 방법론이되며, 매 시간 t에 단어 k의 HMM 상태 i에 도달하는 optimal path의 누적 점수들과 앞의 단어 j로부터 천이후 단어 k의 초기 상태에 도달되는 optimal path 의 누적 점수들을 계산하여 path updating을 수행한다. 이 path updating에서 고려해야 할 변수들은 아래와 같다. 변수명은 구현된 로보트 구동 시스템에 사용한 것이다.

sum : 시간 t에서 복장 벡터의 local likelihood 점수
tempmax : 바로 전 단어의 끝으로부터 다음단어 jj의 상태 k로 천이가 일어날 경우 best path의 누적 점수
likmax : 단어 jj의 바로 전 상태에서 같은 단어 jj의 다음 상태 k로 천이가 일어날 경우 best path의 누적 점수
lik[jj][k] : 시간 t에 단어 jj의 상태 k에 도달하는 best path 의 누적 점수
likprev[jj][k] : 시간 $t-1$에 단어 jj의 상태 k에 도달하는 best path의 누적 점수
LIKprev[ii] : 시간 $t-1$에 단어 ii의 최종 상태에 도달하는 best path 의 누적 점수
LIK[ii] : 시간 t에 단어 ii의 최종 상태에 도달하는 best path의 누적 점수
wtrack[t][jj] : 시간 t에 가장 좋은 likelihood 점수의 전 단어
elapse[ii][k] : 단어 jj에 들어온 후 시간 t까지의 duration 길이
wlength[t][ii] : 시간 t에 끝나는 단어 jj의 총 duration

위의 변수들에는 다음의 관계가 있다.

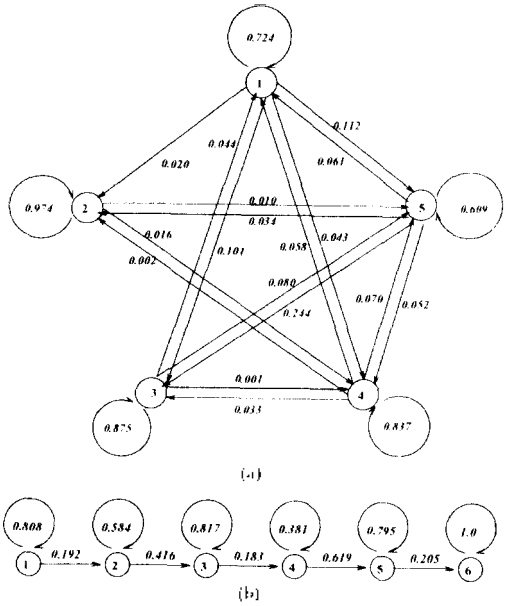

그림 3: EM 알고리즘에 의해 추정된 HMM의 상태 천이 (a)silence HMM의 상태 천이, (b)단어 '어제' HMM의 상태 천이

$$\mathbf{tempmax} = \max_{ii}\{\mathbf{LIKprev}[ii] + A[ii][jj] + init[jj][k]\}$$

$$\mathbf{wtrack}[t][jj] = \arg\max_{ii}\{\mathbf{LIKprev}[ii] + A[ii][jj] + init[jj][k]\}$$

$$\mathbf{likmax} = \max_{jj \in \{\text{states of word } jj\}}\{\mathbf{likprev}[jj][j] + a[jj][j][k]\}$$

$$\mathbf{lik}[jj][k] = \max(\mathbf{tempmax}, \mathbf{likmax}) + \mathbf{sum}$$

위에서 $A[ii][jj]$는 단어 ii에서 단어 jj로의 천이 모수의 로그값이며, $init[jj][k]$는 단어 jj의 초기 상태 k의 확률의 로그값이다. $a[jj][j][k]$는 단어 jj HMM의 상태 j에서 상태 k로의 천이 확률의 로그 값이다.

탐색 알고리즘은 다음과 같이 구현되었다.

1. DP 탐색 알고리즘의 network 상의 초기조건을 부여한다. 관련 변수들의 초기화
2. 매 시간 t,($t = 1, \cdots,$ 프레임 길이T) 에 누적된 likelihood 점수들과, optimal path를 역추적하는데 필요한 관련 변수들을 update한다.
   각 단어 k 에 대하여 optimal path 의 update 는 다음 중 한가지로 된다.
   - 단어 k 모델 안에서 :
     local distance(observation과 상태 천이와 로그 확률 값)의 그 프레임에 대한 계산
     시간 t에 단어 k의 각 상태에 도달하는 best path의 누적 점수 update
     시간 t에 단어 k가 끝나는 경우에 대한 누적 점수 update
     duration 변수의 증가
   - 단어들의 grammar network 레벨에서 :
     전 단어로 부터의 탈출과 다음 가능한 단어의 초기 상태로의 천이
     maximum 누적 점수를 주는 가장 좋은 전 단어를 찾는다 (optimal 단어열 역추적을 위한 정보 저장)
     시간 t에 단어 k의 초기 상태를 시작하는 best path의 누적 점수 update
     duration 변수를 1로 setting
3. 시간 T에서의 optimal 최종 단어로부터 추적 변수를 사용 역추적하여 가장 좋은 likelihood 점수의 단어열을 얻는다.

5 실험 및 결과

A	본체, 팔목, 어깨, 팔꿈치, 팔목, 손가락
B	왼쪽, 오른쪽, 좌, 우, 위, 아래, 아래로
C	십십도, 구십도, 백팔십도
D	올려, 내려, 돌려, 멈춰, 단아
G	garbage 단어(를,을,어,아,주세요,바요,음 등)

표 1: grammar network을 위한 단어그룹 정의

로보트 제어용 학습 데이타는 그림 2를 통해 표 2와 같은 개념적 문 단위를 통해 수집하였으며, 각 단어그룹은 표 1과 같이 정의해서 적용하고 있다. 2.5초 이내에 추어진 문장을 자연스럽게 세번 발성하도록 유도함과 동시에 숨소리, 감탄사 등의 비핵심어도 문장내에 포함시켜서 녹음하였다.

학습과 인식실험을 위해서 시내전화를 대상으로 35명분의 데이타를 수집하였으며, 수동으로 음성구간을 추출, 이중 15명의 데이타를 학습에, 10명분의 데이타는 실험용으로 활용하였다. 입력음성신호는 대역통과필터, 종폭, 음성 구간 검출을 거쳐 LPC 분석을 수행한다. LPC 계수에 대해 켑스트럼과 차분 켑스트럼을 구하며, 켑스트럼(12차), 차분 켑스트럼(12차), 파워 및 차분파워(2차)를 하나의 벡터내에 연속으로 붙여서 만들어진 26차의 특징벡터를 각 음성구간 프레임마다 도출해서 그대로 HMM의 모수추정과 인식에 사용했다.

각 단어별로 학습을 수행 연속 HMM의 모수를 산정하였고 그림 2의 문법 제약 조건에 따라 연결 단어 인식을 위한 FSN을 구성하였다. 위에서 설명된 단어와 단어 사이의 천이 모수는 deterministic하게, 즉 천이 가능의 경우 1을 불가능의 경우 0을 주는 방식을 사용하였다. 초기 단어에 대한 확률은 초기 silence와 garbage나 A군에 속한 핵심단어들에게 양수로 주었다.

단어별 학습에 있어, 초기 HMM 모수의 추정치는 uniform 분할과 k-means clustering 알고리즘을 이용하여 구하였다. 즉, 매 학습 utterance 데이타를 해당 단어의 상태 수로 uniform하게 분할하여 각 상태에 속하는 특징벡터들의 집합을 형성하고 k-means clustering 알고리즘으로 clustering하여 observation 확률 밀도 함수의 모수들(cluster의 sample mean vector와 covariance matrix)에 대한 초기값을 정하였다. 초기상태 확률은, ergodic 형 모델의 경우 단어의 모든 상태에 uniform하게 주었으며 우

향 HMM의 경우 단어의 처음 상태에 1을 주었다. 마찬가지로 천이 확률도 ergodic형 모델의 경우 모든 상태에서 모든 상태로의 천이 확률을 uniform하게 주었으며 우향 HMM의 경우 self-transition과 바로 다음 상태로의 천이 확률에 대해 적당한 양수를 주었다. N은 핵심단어의 경우에 음소의 갯수 더하기 3 ~ 5

- [콘센|팔목] {를|을} [왼쪽|오른쪽] {으로}
  {<각도>} {돌려} {봐} {주세요}
- [어깨|팔꿈치|팔목] {를|을} [위로|아래로]
  {<각도>} {올려|내려} {봐} {주세요}
- 손가락 {을} [펼려|펼어|당아] {봐} {주세요}
- [전원데] 위치 {로} {가}
- [처소|집어|놓아|조금더]
- 커피 {한잔} [따라|따르시오]
- <각도> ::= [30도|90도|180도]

표 2 학습 및 실험용 예문

을 주었으며, silence 및 background 잡음에는 5개, garbage HMM에는 10개 정도를 주었다. mixture의 수 M은 2 ~ 3 개 주었다. 학습 데이타의 양의 부족으로 M이 클 경우 covariance matrix의 determinant가 0에 가까워지는 경우가 많았다. 이는, covariance matrix는 diagonal한 행을 사용하였으므로, 형성 가능 mixture의 수가 많아지면서 특징 벡터의 component 들 중 거의 같은 부분을 가진 작은 수의 특징 벡터들이 하나의 mixture component의 재추정에 영향을 미침으로써 발생한다고 이해된다. EM 알고리즘 학습의 반복 수는 최대 10번으로 했다. 대부분의 경우 10번 이전에 주요한 우도 점수의 수렴이 이루어졌다.

인식의 출력은 역추적에 의해 얻어진 optimal 단어열, 즉 단어들의 index 열과 각 단어들의 duration의 열이다. 인식 속도는 실험 문장의 발음 길이에 따라 변하지만 보통 속도의 발음에 대해 각도 부분을 생략하여 구성할 경우 2 ~ 3 초, 각도 부분을 포함시킬 경우 7 ~ 13 초 걸렸다. 완성된 인식 시스템에 대한 인식 실험 결과 학습 화자에 대한 제어 명령 인식률은 95%이상을 나타내고 있으며, 비 학습 화자에 대한 인식율은 90%이상이다.

6 결 론

본 논문에서는 음성인식의 실용화를 위한 로보트 제어용 인터페이스 시스템에 대해 기술하였다. 시스템 구성에 있어서는 PC-486 시스템에 TMS320C31에 기반한 Elf-31 신호처리보드를 적용했는데, 특히 본 시스템에서는 4장까지의 신호처리보드를 동시 다중처리할 수 있도록 시스템을 구축하였으며, 인식면에서 연속 분포 HMM에 의한 keyword spotting 기법을 통해 자연스런 음성 명령을 통한 로보트 제어를 실현할 수 있었다.

비 핵심단어들에 대해서 ergodic형의 HMM에 의한 학습은 다수의 단어(비핵심 단어)들을 한데 묶어서 하나의 범주를 만들어 그에 대한 전체적인 통계적 특성 즉 관심없는 단어들에 대한 대체적 성질 을 형성하는 데 성공적이라고 할 수 있겠다. 이는 구현된 시스템에 대한 실시간 인식 실험에서 비핵심 단어들의 사용에 대해 매우 flexible한 인식 성능을 관찰함으로써 확인된다. 중간 중간의 불규칙한 duration의 감탄사의 삽입에 대하여, 또 심지어 본 실험 문장의 시작이전이나 끝난 이후에 비 핵심 단어들로 구성된 다른 어떤 문장을 삽입하는 경우에도, 시스템의 올바른 핵심 단어 인식 결과를 낼 수 있었다. 물론 각각의 garbage class로 합쳐져서 모델링된 비핵심 단어들의 acoustic 특징이 핵심 단어들의 그것과 많이 다를 경우에 effective한 방법론이겠다.

참고 문헌

[1] F. Jelinek, "A Fast Sequential Decoding Algorithm using a Stack", *IBM J Res. Develop.*, vol. 13, pp. 675-685, Nov. 1969.

[2] C. S. Myers and L.R. Rabiner, "A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition", *IEEE Trans. on ASSP*, vol. 29, pp. 284-297, Apr. 1981.

[3] R. Nakatsu, "Auser: An Application of Speech Technology to the Japanese Banking Industry", *IEEE Computer*, pp. 43-48, Aug. 1990.

[4] H. Ney, "The Use of a One-stage Dynamic Programming Algorithm for Connected Word Recognition", *IEEE Trans. on ASSP*, vol. 32, pp. 263-271, Apr. 1984.

[5] Lawrence Rabiner and Biing-Hwang Juang (1993), *Fundamentals of Speech Recognition*, Ch 7-8, Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

[6] L. R. Rabiner and C. H. Lee, "A Frame-Synchronous Network Search Algorithm for Connected Word Recognition", *IEEE Trans. on ASSP*, vol. 37, pp. 1649-1658, Nov. 1989

[7] H. Sakoe, "Two Level DP Matching - A Dynamic Programming Based Pattern Matching Algorithm for Connected Word Recognition", *IEEE Trans. on ASSP*, vol. 27, pp. 588-595, Dec. 1979.

[8] Y. Takebayashi *et al.*, "Keyword-spotting in Noisy Continuous Speech Using Word Pattern Vector Subtraction and Noise Immunity Learning", *ICASSP 92*, pp. II-85-88, 1992.

[9] J. G. Wilpon and L. Rabiner, "Automatic Recognition of Keyword in Unconstrained Speech Using Hidden Markov Models", *ICASSP 90*, pp. 1870-1878, 1990.