

지능형 홈네트워크 시스템을 위한 화자독립 음성인식시스템 개발에 관한 연구

종신회원 이 호 응*, 정회원 정 희 석**

A Study on a Speak-Independent Continuous Speech Recognition System for Intelligent Home Network System

Ho-woong Lee* *Lifelong Member*, Hee-suk Jeong** *Regular Member*

요 약

본 논문에서는 지능형 홈네트워크의 음성제어를 위한 화자독립 연속음성인식엔진을 개발하였다. 또한 자연스런 음성명령에 대한 인식을 위해 핵심어 기반의 자연스런 연속어휘에 대한 대화형 시나리오를 작성하였고, 핵심어기반의 인식엔진 및 데이터베이스를 구축하여 인식엔진의 성능을 최적화하였다.

Key Words : Intelligent Home Network System, Speech Recognition, Speech Independent, HMM

ABSTRACT

This study aims to develop a continuous speaker-independent speech recognition system for the speech control of intelligent home-network. This study suggests a conversational scenario of continuous natural vocabulary based upon keywords for recognition on natural speech command, and a way of optimizing the recognition system by constructing a recognition system and database based upon keywords.

I. 서 론

최근 몇 년간 Environmental Robustness 분야는 음성 인식 연구 분야 중에 가장 주목받고 있으며, 많은 연구소에서 음성의 음향학적인 특징, 음성 특징들의 필터링에 기반을 둔 접근과 다른 알고리즘들로 인식 시스템의 정확성을 증대시키기 위해 연구되고 있다^{[1][2][3]}. 음성인식 전반적으로는 현재의 통계적 방법을 기반으로 실재의 대량의 음성데이터에 기초를 둔 일상 언어의 언어모델을 구축하는 것, 다수화자의 음성데이터에 기반하여 개인자의 모델을 구축하여 이에 의한 다수 화자의 음성에의 적응화 알고리즘을 개발하는 것, 여러 종류의 잡음, 왜곡에 자동적으로 적응되는 방법을 확립하는 것 등이 중

요한 기술적 과제로 될 것이다. 음성인식 기술을 적용한 홈네트워크 시스템은 잡음제거 환경에서 음성 인식 성능, 원거리 음성인식 성능, 비대상어휘 제거 성능 등이 중요한 평가 지표가 될 것으로 보인다^{[4][5]}. 현재 대부분의 홈네트워크 전문회사는 연간 30 만호에 달하는 신규 아파트 및 주택을 대상으로 과거의 단순한 홈오토메이션시장에서 각종 가정 내 전등제어 및 냉난방공조시스템, 가전제품의 제어를 통합하고 초고속 인터넷을 이용하여 각종 정보를 손쉽게 접근할 수 있고 집밖에서도 집안내의 시스템을 제어할 수 있는 지능형 홈네트워크 시스템으로 진화시키고 있다. 따라서, 기존의 단순한 기능에서 다기능화된 시스템을 효율적이고 간편하게 접근할만한 새로운 인터페이스 기술이 요구되고 있으며

* 동원대학 정보통신과(hwlee@tongwon.ac.kr), ** (주)파워보이스 대표이사
논문번호 : 07103-1218, 접수일자 : 2007년 12월 18일

많은 건설사들과 홈네트워크 전문회사들은 이러한 인터페이스를 제공하기 위해 음성인식시스템에 대한 요구가 급증하고 있는 실정이다. 본 논문에서는 화자독립 핵심어기반 연속음성인식을 적용한 지능형 홈네트워크 시스템의 개발에 대한 내용을 제안하고자 한다.

II. 본 론

2.1 음성인식 시스템 개요

인간이 자기의 생각을 다른 사람에게 표현할 수 있는 수단은 여러 가지의 형태를 보인다. 그중 음성은 인간이 사용하는 가장 기본적인 의사소통을 위한 수단이며, 편리함과 경제성의 측면에서 다른 방법에 비하여 우수한 특성을 가진다. 인간에 의한 음성처리는 크게 음성생성(Speech production)과 음성인지(Speech perception)의 두 가지 측면으로 나누어 볼 수 있다. 음성생성은 발화자(Speaker)가 의도한 바를 전달하기 위한 일련의 과정이고, 음성인지는 상대 발화자에 의해서 발생된 음성으로부터 발화내용을 인식하는 과정을 말한다. 이러한 음성의 두 가지 측면과 관련된 연구들이 각기 개별적으로 이루어져 왔으며 언어학, 음성학, 음운학, 생리학, 해부학 등 다양한 학문적인 배경 하에 진행되어 왔다. 이러한 결과들이 신호처리기술, 기억장치의 대용량화, 그리고 고속의 컴퓨터 처리기술의 발달 등 급격한 기술의 발전으로 인해 단순히 실험적인 결과가 아닌 실용적인 측면에서 그러한 결과들을 활용하는 연구가 활발히 진행되어 왔으며, 계산이론적인 측면에서 음성처리와 관련된 다양한 연구들이 이루어지게 되었다⁶⁾.

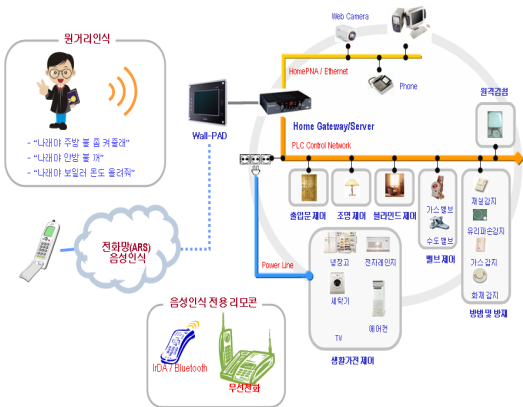


그림 1. 지능형 홈네트워크시스템을 위한 음성인식시스템

2.2 Model-Based HMM(Hidden Markov Model) 알고리즘³⁾

음성인식을 수행하기 위하여 여러 가지의 접근방법이 존재한다. 그 중 패턴매칭 접근방법은 소규모의 고립단어로부터 대규모의 연속음성에서까지 우수한 인식성능을 보이는 접근방법으로서 크게 Template-Based 방법인 Dynamic Time Warping 방법과 Model-Based 방법인 Hidden Markov Model로 분류할 수 있다. 본 논문에서는 비교적 인식율이 우수한 것으로 밝혀진 HMM기반의 음성인식엔진을 설계하였고, 잡음환경에 강인한 Model-based HMM 알고리즘을 개발하여 적용하였다. HMM은 음성신호의 스펙트럼 변화 및 시간 변화를 동시에 모델링할 수 있으며, 이를 위하여 유한개의 상태와 상태전이들을 사용한다. HMM의 유용성은 음성 생성 과정을 정확히 모델링 할 수 있는 것이 아니라, 오히려 주어진 데이터를 사용하여 파라미터를 추정하고 새로이 입력된 음성에 대하여 가장 적합한 모델을 찾는 데 있다. 즉, 훈련과 인식을 위한 효율적인 알고리즘이 존재한다는 것이다. 모델의 파라미터를 추정하기 위하여 카테고리 정보가 있는 음성 데이터 베이스를 사용하며, 각 모델을 위한 충분한 데이터가 있을 경우 실제 음성에 존재하는 다양성을 잘 표현할 수 있는 강인한 모델링이 가능하다. 중규모 혹은 대어휘 시스템을 위하여 각 모델을 정확하게 추정할 수 있을 만큼의 충분한 데이터를 확보하기 위해서는 단어 단위 대신 음소와 같은 Sub-word 단위의 모델이 주로 사용된다. 훈련용 음성데이터는 일반적으로 완전한 문장 형태로 주어지기 때문에 음소 단위 모델을 문장 단위의 데이터에 매칭시키기 위해서는 문장을 음소 단위로 정확하게 분할하는 것보다 전체 문장을 위하여 음소모델을 연결한 복합 모델을 생성하여야 한다. 발생된 음성 데이터를 음소 단위로 정확하게 분할하는 것은 어렵기 때문에 전자보다는 후자가 주로 사용되고 있다. 복합 모델을 구성하기 위해서는 각 발성에 대한 음소 수준의 트랜스크립션(Transcription)이 필요하며, 일반적으로 사전을 사용하여 단어를 음소 수준으로 변환한다.

HMM은 일련의 연속된 상태(State)들로부터 이산 신호를 생성하는 확률 과정모델이다. 모델은 전이확률(Transition probability)에 따라 상태를 바꾸면 특정 상태는 그 상태의 출력확률(Output probability distribution)에 따라 하나의 관측(Observation)을 발생시킨다.

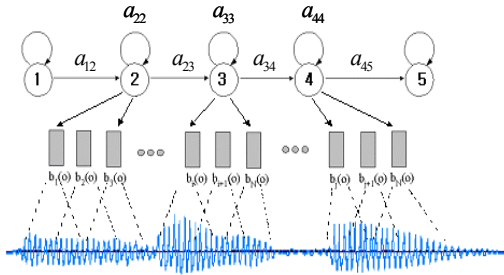


그림 2. 음성신호의 HMM모델

음성을 모델링하기 위하여 여러 가지 형태의 HMM이 사용되고 있지만 그림 2와 같은 간단한 구조의 left-to-right 모델이 많이 사용되고 있다. 상태 2, 3, 4번은 출력이 있는 상태들이며 1번과 5번 상태는 출력은 없고 단지 모델의 연결을 도와주는 기능을 수행한다. o_t 는 음성 벡터 혹은 관찰을 나타내고 a_{ij} 와 $b_j(o_t)$ 는 각각 전이확률과 출력확률 분포함수를 의미한다. 전이확률 a_{ij} 는 상태 i 에 있던 모델이 상태 j 로 상태를 변화시킬 조건부 확률로서 다음 식 (1)과 같다.

$$a_{ij} = P(x(t+1) = j | x(t) = i) \quad (1)$$

$x(t)$: 시간 t일때 모델의 상태

HMM이 출력을 생성하는 구조로 사용될 뿐만 아니라 $o(1), \dots, o(T)$ 의 프레임으로 구성되는 신호 o 가 특정한 상태열 $X = x(1), \dots, x(T)$ 에 의하여 생성될 우도(likelihood)를 계산하기 위해서도 사용된다. 이것은 각 상태에서 특정 관측이 발생할 우도와 전이확률로부터 계산되는 상태열에 대한 확률을 곱함으로써 다음 식 (2)와 같이 구해진다.

$$P(O, X) = a_{12}b_2(o_1)a_{23}b_3(o_2)a_{34}b_4(o_3)a_{45}b_5(o_4)a_{55}b_5(o_5) \quad (2)$$

전이확률은 상수로 가정하여 시간에 따라 변하지 않는다. 전체 상태의 개수를 N이라 했을 때, 모든 초기 상태들 $i = 1, 2, \dots, N-1$ 는 다음 식 (3)의 조건을 만족해야한다.

$$\sum_{j=2}^N a_{ij} = 1.0 \quad (3)$$

출력확률 분포함수 $b_j(o_t)$ 는 상태 j에 의하여 발생될 관측들의 분포를 나타낸다. 이것은 상태 j가 관

측 o_t 를 생성할 확률(Likelihood: 연속 출력 분포인 경우)을 의미한다. HMM에서 출력이 있는 모든 상태들은 그 상태에서 발생하는 관측(Observation)들의 확률을 결정하기 위하여 출력확률 밀도함수를 가진다.

밀도함수로는 이산(Discrete), 연속(Continuous), 반연속(Semi-continuous) 혹은 혼합공유(Tied-mixture) 분포가 사용되고 있으며, 비교연구에 의하면 충분한 훈련 데이터만 주어지면 연속 분포모델을 사용한 시스템이 계산량은 많지만 인식 성능은 가장 우수한 것으로 알려져 있다.

출력확률분포는 서로 다른 소리들을 충분히 구분할 수 있으면서 동시에 음성속에 내재된 여러 가지 다양성들을 소화할 수 있을 만큼 강인해야 한다. 이산분포를 사용하는 경우 관측들은 코드북을 사용하여 하나의 심볼로 양자화 되고, 각 상태는 그 상태에 의하여 각각의 심볼이 생성될 확률을 나타내는 하나의 이산분포를 가지게 된다. 관측들 내에서의 변화를 더 잘 표현하기 위하여 복수 코드북들이 사용되는 경우가 많다. 반연속분포는 이산분포의 속도와 연속분포의 정확성을 함께 살리려는 시도이다. 이 분포는 하나의 가우시안 집합을 공유하면서 각 상태별로 다른 가중치 집합을 가지며 다음 식 (4), (5)와 같다.

$$b_m(o_t) = N(o_t | \mu_m, \Sigma_m) \quad (4)$$

$$= \frac{1}{(2\pi)^n |\Sigma_m|} \exp\left\{-\frac{1}{2}(o_t - \mu_m)' \Sigma_m^{-1} (o_t - \mu_m)\right\}$$

$$b_j(o_t) = \sum_{m=1}^M c_{jm} b_m(o_t) \quad (5)$$

여기서 c_{jm} 은 평균 μ_{jm} 과 공분산 Σ_m 을 가지는 m 번째 공유 가우시안에 대한 이 상태의 가중치를 의미한다. 연속 출력 확률 분포인 경우에 단일 혹은 혼합가우시안 확률 밀도함수는 다음과 같다.

훈련과 인식 중에 혼합가우시안 확률을 계산하는데 많은 시간이 소요되며 특히 전체 공분산 행렬이 사용될 경우에는 더욱 심각해진다. 계산시간과 메모리량을 줄이기 위하여 공분산의 대각 행렬 원소들만이 주로 사용되고 있다. 적은 숫자의 혼합과 전체 공분산을 사용하는 것보다 더 많은 혼합과 대각행렬 공분산을 사용하는 것이 더 나은 접근 방법이라는 사실도 알려져 있다. 대각행렬 공분산을 사용하

기 위해서는 특징 벡터의 각 원소가 통계적으로 서로 독립이 되어야 한다. 특징벡터에 대한 켈스트럴 계수(Cepstral coefficient)를 구함으로써 벡터 내의 원소들 간에 독립을 거의 보장할 수 있다. 그러나, Model-based HMM을 통해 음성인식시스템을 구현 하는데 있어서는 다음과 같은 확률계산상의 문제와 모델추정을 위한 학습과정, 인식과정에 대한 다음과 같은 문제점을 해결하여야 한다. 계산수행을 위한 문제(Evaluation problem)는 관측열 특징벡터 x 가 들어왔을 때의 확률이 어떻게 계산되는지의 문제로써 전향알고리즘(Forward algorithm)과 후향알고리즘(Backward algorithm)을 이용하여 해결될 수 있다. 은닉 상태열을 찾는 문제에서는 관측열 $O = \{o_1, o_2, \dots, o_T\}$ 과 모델 $\lambda = (A, B, \pi)$ 가 주어졌을 때, 가장 최적의 상태열 $Q = \{q_1, q_2, \dots, q_T\}$ 를 찾기 위해 Viterbi 알고리즘을 이용한다.

HMM에 기반한 시스템에서 인식이란 다음 식 (6)에서와 같이 알려지지 않은 데이터 열 O 에 가장 적합한 모델 혹은 복합 모델 $\bar{\pi}$ 를 선택하는 것이다.

$$\bar{X} = \max_x \Pr(X|O) \quad (6)$$

여기서, Bayes-Rule을 사용하면 다음 식 (7), (8)이 성립된다.

$$\Pr(X|O) = \frac{\Pr(O|X)\Pr(X)}{\Pr(O)} \quad (7)$$

$$\bar{X} = \max_x \Pr(O|X)\Pr(X) \quad (8)$$

음성인식을 위하여 HMM이 성공적으로 사용되고 있는 중요한 이유중의 하나가 \bar{X} 를 계산할 수 있는 효율적인 알고리즘이 존재한다는 것이다. $\Pr(O|X)$ 를 계산하기 위하여 전향 확률(Forward probability)을 구하는 방법을 사용할 수도 있지만 연속 음성인식을 위해 효과적으로 사용될 수 있는 최대우도 상태열에 기반한 방법이 주로 사용되고 있다. 음성벡터 o_1, \dots, o_t 를 생성하면서 시간 t 에서 상태 j 에 있을 최대확률 $\varnothing_j(t)$ 는 다음 식 (9)와 같은 순환식에 의하여 계산될 수 있다.

$$\varnothing_j = \max_x \{\varnothing_i(t-1)a_{ij}\}b_j(o_t) \quad (9)$$

수치문제 발생을 피하기 위하여 로그확률을 사용

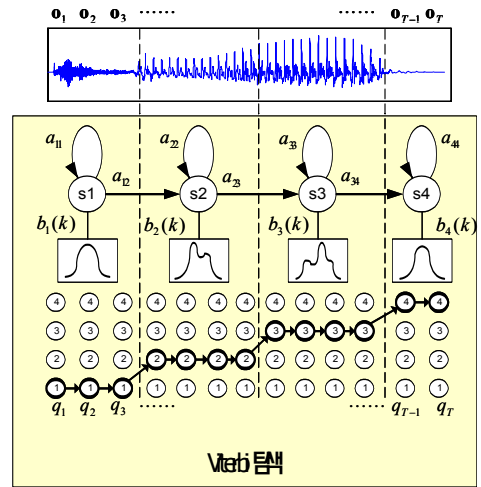


그림 3. 비터비 알고리즘

하면 다음 식 (10)과 같이 표현되며 이것이 Viterbi 알고리즘의 근간이 되는 식이다.

$$\phi_{j(t)} = \max_i \{\phi_i(t-1) + \log(a_{ij}) + \log(b_j(o_t))\} \quad (10)$$

Viterbi 알고리즘은 고립 단어 인식 시스템에서는 그대로 사용되지만 연속음성인식을 위해서는 토큰 전달알고리즘으로 확장 구현되어 많이 사용된다. 현재까지의 로그 확률값 $\phi_{j(t)}$ 를 가지고 있는 토큰을 다음상태로 전달하며 모든 상태에서 토큰들을 검사하여 가장 높은 확률값을 가지고 있는 토큰 외에는 기각시키는 방법으로 현재 선택된 상태의 토큰은 새로운 확률값으로 갱신시키고 추후 상태열의 복원을 위하여 현재 선택된 상태는 기록된다. 이 방법은 모델 혹은 단어 수준으로 그대로 확장시킬 수 있다.

HMM에서의 학습문제에서는 관측열의 확률 $P(O|\lambda)$ 를 극대화시키기 위해 모델 매개변수 $\lambda = (A, B, \pi)$ 를 재추정하기 위해 Baum-Welch 알고리즘을 통해 해결할 수 있다.

$\bar{\pi}_i$: 시간 $t=1$ 에서 상태 i 에 있을 확률

$\bar{a}_{ij} = \frac{i\text{상태로부터 } j\text{상태로 천이할 확률}}{i\text{상태로부터 천이할 확률}}$

$\bar{b}_j(k) = \frac{\text{관측되는 심볼이 } v_k\text{이고 상태 } j\text{에서 천이할 확률}}{\text{상태 } j\text{에 있을 확률}}$

$$(11)$$

여기서, 새 파라미터, $\bar{\lambda} = (\bar{\pi}, \bar{A}, \bar{B})$ 는 $P(O|\bar{\lambda}) > P(O|\lambda)$ 일 때 근사화된다.

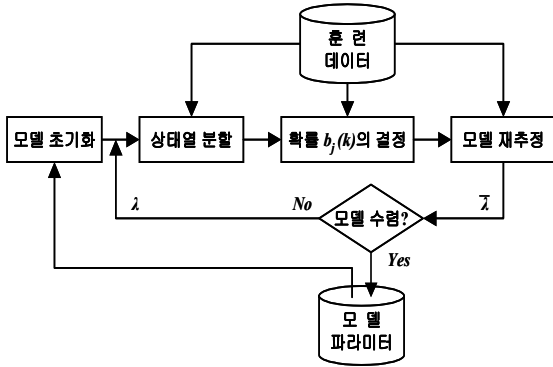


그림 4. Baum-Welch 알고리즘 흐름도

Ⅲ. 실험환경 및 결과

3.1 실험환경

본 논문에서 사용된 실험환경은 다음과 같다. 화자독립 연속음성인식을 위한 명령어 셀은 기본 PLU를 분석하기위한 표준 음성데이터베이스를 이용하였고, 홈네트워크에서 흔히 이용가능한 단어군을 구성하여 서울, 경기지역의 표준말과 강원도, 충청도, 전라도, 경상도, 제주도의 각 지역별 음성데이터베이스를 구축하였다. 또한, 화자독립형 음성데이터베이스 구축을 위해 10대와 20~30대, 40대 이상의 성별, 연령별 음성을 취득하고 이를 기반으로 화자독립형 음성데이터베이스를 구축하였다. 본 논문에 사용된 홈네트워크용 연속음성인식기 인식어휘의 명령어 셀에 대해서는 아래와 같이 기술하였다.

분류	명령어		
	키워드	제어대상	제어명령
조명제어	지니 나래야	1. 전체 2. 안방 3. 거실 4. 작은방 5. 서재 6. 베란다 7. 주방 8. 손님방	불켜
			불꺼
			불 밝게
			불 어둡게
			조명 켜
			조명 꺼
			조명 밝게
			조명 어둡게
난방제어	지니 나래야	1. 전체 2. 안방 3. 거실 4. 작은방	보일러 켜
			보일러 꺼
			보일러 온도 올려
			보일러 온도 내려
냉방제어	지니 나래야	1. 전체 2. 안방	에어컨 켜
			에어컨 온도 올려
			에어컨 온도 내려

분류	명령어	
	키워드	제어명령
가스밸브	지니 나래야	가스밸브 열어
		가스밸브 닫아
블라인드	지니 나래야	블라인드(커튼) 열어
		블라인드(커튼) 닫아

3.2 실험결과

1) 음성인식률

단위	결과	평가방법
%	98%	집안환경/40dB잡음환경/3m거리
	95%	집안환경/45dB잡음환경/3m거리
	90%	집안환경/50dB잡음환경/3m거리

2) 거절률

단위	결과	평가방법
%	99%	집안환경/음악,TV소음환경에서의 오동작실험(24시간, 28,800건의 이벤트 중 1건 발생)

3) 원거리인식

단위	결과	평가방법
m	98%	테스트패널에서 3m 거리
	94%	테스트패널에서 4m 거리
	92%	테스트패널에서 5m 거리

Ⅳ. 결 론

본 논문은 “지능형 홈네트워크시스템을 위한 음성인식시스템 개발”을 목표로 실생활에 적용가능한 원거리 음성인식모듈을 개발하는 것으로서 화자독립 핵심어기반 연속음성인식엔진 개발과 원거리인식용 화자독립형 음성데이터베이스 개발 등에 관해 연구하였다. 잡음에 강인한 원거리 음성인식 모듈은 3~5m의 거리에서도 원활한 음성인식을 위해 잡음 제거 및 원거리 인식용 데이터베이스 구축에 많은 진전을 보였으며, 이를 통해 3m 이상의 원거리에서 잡음환경 40~50dB의 실생활 환경에서 95%라는 높은 인식율을 보이게 되었다. 향후 연구과제로 음성인식 시스템의 상용화에서 가장 걸림돌이 되어왔던 미등록어에 대한 연구와 음성인식전용 DSP모듈 개발에 관한 연구가 지속되길 기대한다.

참 고 문 헌

[1] B. Juang, “Speech Recognition in Adverse

Environments,” *Computer Speech and Language*, Vol.5, pp.275-294, 1991.

[2] M. F. Gales, “Model-Based Techniques for Noise Robust Speech Recognition,” *Ph.D dissertation, University of Cambridge*, Sept. 1995.

[3] L. Rabiner, Biing-Hwang Juang, *Fundamentals of Speech Recognition*, Prentice-Hall International, Inc., 1993.

[4] Takatoshi Jitsuhiro, Satoshi Takatoshi, Kiyoaki Aikawa, “Rejection of Out-of-Vocabulary Words using Phoneme Confidence Likelihood”, *ICSSP*, pp.217-220, 1998.

[5] Alexandros S. Manos, Victor W. Zue, “A study on out-of-vocabulary word modeling for a segment-based keyword spotting system”, 1996.

[6] Daniel Jurafsky & James H. Martin, “Speech and Language Processing”, *Prentice Hall*, 2000.

이 호 응 (Ho-Woong Lee)
2007년 12월호 참조

중신회원

정 희 석 (Hee-suk Jeong)

정회원



1996년 8월 광운대학교 전자 통신공학과 졸업(공학사)
1998년 8월 광운대학교 대학원 전자통신공학과 졸업(공학석사)
2005년 2월 광운대학교 대학원 전자통신공학과 졸업(공학박사)
2003년 3월~현재 (주)파워보이스 대표이사

<관심분야> 음성인식, 화자인식, 패턴인식, 적응신호처리