논문 2011-3-24

모델적응 HMM을 이용한 모바일환경에서의 음성인식에 관한 연구

A study on Voice Recognition using Model Adaptation HMM for Mobile Environment

안종영*, 김상범**, 김수훈***, 허강인****

Jong-Young Ahn, Sang-Bum Kim, Su-Hoon Kim, Kang-In Hur

요 약 본 논문에서는 모바일 환경에서의 음성인식 개선에 관한 내용으로 기존의 HMM에서 특징보상기법을을 적용한 방식으로 예측잡음이 아닌 실제 오염된 데이터를 적용하여 인식모델을 잡음상황에 맞도록 적응시키는 모델적응 HMM을 사용하였다. 음성인식 시 기존의 방법에서는 주변노이즈를 고려하지 않은 참조패턴을 사용하였으나 본 연구에서는 주변노이즈를 고려한 참조패턴을 생성하여 인식률을 향상 시키는 방법으로 모바일 환경에서의 음성 인식률을 향상 시켰다.

Abstract In this paper, we propose the MA(Model Adaption) HMM that to use speech enhancement and feature compensation. Normally voice reference data is not consider for real noise data. This method is not to use estimated noise but we use real life environment noise data. And we applied this contaminated data for recognition reference model that suitable for noise environment. MAHMM is combined with surround noise when generating reference patten. We improved voice recognition rate at mobile environment to use MAHMM.

Key Words : Voice Recognition, HMM, Noise Cancel

Ⅰ. 서 론

컴퓨터 및 정보통신 기술의 급속한 발전으로 음성인식 기술은 중요한 연구과제가 되고 있다. 최근 음성인식 기술은 상용화 가능한 수준까지 도달 하였고 미국, 일본 등지에서는 이미 연속 음성을 인식할 수 있는 상업용 제품들이 생산 판매되고 있으며 이와 관련된 많은 연구가 진행되고 있다. 그러나 아직도 실제 환경에서는 같은 환경이라 할지라도 그 인식률은 많은 차를 나타낸다. 특히, 음성인식에서 가장 큰 변수 요인이 바로 주변 잡음이라 할 수 있다. 그 예로 모바일을 이용한 음성인식 즉, 핸드폰 또는 PDA나 차량이동시의 음성인식에 있어서 주변 환경 잡음에 의한 영향으로 음성 인식률이 감소하게 된다. 따라서 음성인식시장을 활성화하는데 가장 큰 문제점으로는 잡음이라고 할 수 있다.[1] HMM은 음성패턴의 변동을 통계적으로 처리하고 이 통계량을 확률형태의 모델에 반영하여 음성을 인식하는 방법으로서 개인차나 조음 결합 등의 영향에 의한 음성패턴의 변동을 반영하기 쉽고 확률 통계론에 의한 이론적 전개가 용이하며, 음소나 음절단위의 모델을 단어, 문장 등의 단위로 쉽게 확장할 수 있는 장점이 있으며, 신경망은 화자의 개인차 등에

*정회원, 한국폴리텍2대 컴퓨터정보과
**한국폴리텍2대 컴퓨터정보과
***부천대학 모바일통신과
****동아대학교 전자공학과(교신저자)
접수일자 2011.3.28, 수정일자 2011.5.16
게재확정일자 2011.6.10

의한 스펙트럼의 변동을 유니트간의 결합 가중치로서 표현할 수 있고 한 번에 많은 프레임의 데이터를 입력할 수 있는 장점이 있다. 특징보상 기법은 잡음에 의하여 오염된 특징을 깨끗한 음성에서 추출된 특징으로 변환하고 깨끗한 음성으로 학습시킨 인식 모델을 사용하는 기법으로 크게 mathematical model-based 기법과 datadriven기법으로 분류된다. 기존의 모델적응 기법은, 특징보상 방법과는 달리, 입력 특징은 수정하지 않고 대신 인식 모델을 잡음 황에 맞도록 적응시키는 방식이다. 현재 거의 모든 음성인식 시스템에서는, hidden Markov Model(HMM)을 인식모델로 채택하고 있는데 이들 HMM은 많은 양의 오염되지 않은 음성으로 학습을 한 것이다.[2]

본 연구에서는 상기 기법 중 특징보상방법을 적용하여 모델적응에 적용한 방법으로 HMM을 사용하여 음성인식 시 주변노이즈를 고려한 패턴매칭레벨을 분류하여 인식률을 향상 시키는 방법인 기존 HMM방식에서 잡음이 부가된 데이터를 함께 사용 하는 MA(Model Adaptation)HMM을 제안한다.

II. 본 론

1. HMM(Hidden Markov Model)

마코프(Markov)모델은 수학 및 공학에서 널리 사용되는 방법이다. 그러나 모델 자체의 강한 제한성으로 인해 음성인식 등과 같은 복잡한 문제의 응용에 적용되지는 못하였다. 그러던 중 1960년대 말에 L.E.Baum등은 마코프 체인의 개념을 확장하여 각 상태에서 다른 상태로 천이 할 때 일어나는 확률로 만들어진 천이확률 분포(Transition Probability Distribution), 각 상태에서의 관측 확률분포와 초기 분포로 구성된 HMM을 제안하였다.

음성인식에 HMM을 이용할 때 음성은 발성구조의 시간적 변화에 의하여 발생된 신호이므로 프로세서가 한쪽 방향으로만 천이가 가능하도록 제한시킨 Left-to-right 모델을 주로 사용하며 초기상태와 최종상태가 설정되어 1회의 상태천이 마다 심벌을 1개씩 출력한다. 다음에 어느 상태로 천이 하는가 또 그때에 어떤 심벌을 출력하는가는 각각 천이확률과 출력확률에 의해서 통계적으로 결정되어진다.[3]

HMM은 출력심벌에 의해서 상태천이 경로가 하나로 결정되지 않는다는 의미로 비결정성 유한상태 오토마타(Automata)로 정의 할 수 있다. 일반적인 마코프 모델과 다른 점은 출력 심벌 계열이 주어져도 그 상태 계열은 하나의 경로로 결정되지 않는다는 것이다. 그림 1에서 아크 위의 벡터 값은 심벌 $\{a,b\}$ 의 상태천이에 의한 조건부 출력확률을 나타내고 있다. 즉 벡터의 제 1요소가 a의 출력확률이며 제 2요소가 b의 출력확률을 나타내고 있다. 이 경우 관측 심벌계열이 abb인 경우 가능성이 있는 상태천이 계열은 $S_1S_1S_2S_3$와 $S_1S_2S_2S_3$의 두 경우가 있으며 관측할 수 있는 것은 심벌계열 뿐이고 상태 그 자체는 직접 관측되지 않는다. 이러한 의미로 "Hidden" 마코프 모델이라 부른다.

음성인식에 이용되는 HMM을 구성하는데는 3가지 요소가 있는데, 상태수 N, 시간에 따른 상태의 변화를 결정하는 천이확률, 그리고 각 상태에서의 출력확률이 있다. 이러한 요소들에 의해 구성되는 HMM에서는 다음의 8가지 요소가 조합된 $M=(T,N,S,Y,A,B,\pi,F)$로 정의된다.[4]

1) T : 관측열의 길이
2) N : 상태수
3) S : 상태의 유한집합 ; $S=\{s_i\}$
4) Y:출력심벌의집합;
   $Y=\{y_1, y_2, \cdots, y_N\}$
5) A : 상태천이확률의 집합 ; $A=\{a_{ij}\}$;
   a_{ij}는 상태 s_i로부터 상태 s_j 의 천이확률, 단 $\sum_j a_{ij}=1$.
6) B : 출력확률의 집합 ;
   $B=\{b_{ij}(n)\}$; $b_{ij(n)}$는 상태 s_i에서 상태 s_j로 천이할 때 심벌 y_n을 출력하는 확률,

$$\sum_n b_{ij}(n)=1 \quad \text{(이산 HMM)}$$

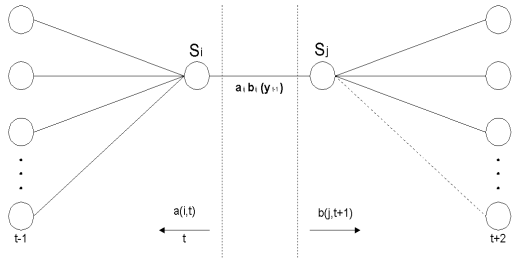
$$\int_{-\infty}^{+\infty} b_{ij}(n)dn=1 \quad \text{(연속HMM)}$$

7) π : 초기상태 확률의 집합 ; $\pi=\{\pi_i\}$ π는 초기상태가 s_i인 확률 단, $\sum_j \pi_j = 1$.

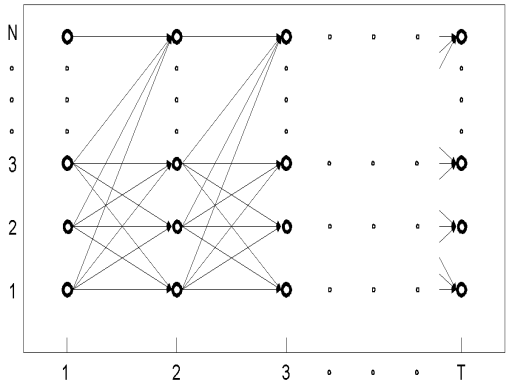
8) F : 최종상태의 집합

학습되어지는 방법은 통상은 Baum-Welch에 의하여 제안된 Forward-Backward 알고리즘을 사용한다.

Forward-Backward 알고리즘은 식(3)과 같이 정의되는 전향확률과 식(7)와 같이 정의되는 후향확률을 이용하여 출력확률 $P(Y \mid M)$을 계산하는 알고리즘으로, 이 방법을 이용하면 계산량과 복잡도를 크게 줄일 수 있다.



(a)



(b)

그림 1. $\alpha_t(i)$와 $\beta_t(i)$ 계산
(a) 전·후향 확률
(b) 전향확률과정

Fig. 1. Calculation of $\alpha_t(i)$ and $\beta_t(i)$
(a) forward and backward probability
(b) calculation procedure of forward probability

$\alpha_t(i)$를 그림 1(a)와 같이 초기상태로부터 $y_1, y_2, \cdots, y_t$를 생성하면서 상태 i에 도달하는 확률 변수로 정의하면, 전향확률은 식(28)을 반복적으로 계산하면 출력확률 $P(Y \mid M)$을 구할 수 있다.

먼저 전향확률 변수 $\alpha_t(i)$를

$$\alpha_t(i) = P(y_1, y_2, \cdots, y_t, i_t = s_i \mid M) \tag{1}$$

로 정의하고 다음 과정을 반복한다.

$$\alpha_1(i) = \pi_i b_{i(y_1)} \quad ; \quad 1 \le i \le N \tag{2}$$

$$t = 1,2,\cdots,T-1 \quad ; \quad 1 \le j \le N$$

$$\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(i) a_{ij}] \cdot b_{j(y_{t+1})} \tag{3}$$

$$P(Y \mid M) = \sum_{i}^{N} \alpha_T(i) \tag{4}$$

그림 1(b)는 $\alpha_t(i)$의 계산순서 개념도를 나타내고 있다.20) 같은 방법으로 $\beta_t(i)$를 최종상태로부터 $y_{t+1}, y_{t+2}, \cdots, y_T$을 생성하면서 상태 i에 도달하는 확률을 후향확률 변수라 하고 $\beta_t(i)$를

$$\beta_t(i) = P(y_{t+1}, y_{t+2}, \cdots, y_T \mid i_t = s_i, M) \tag{5}$$

로 정의할 경우 식(3)을 반복처리 함으로써 구할 수 있다.

$$\beta_T(i) = 1 \quad ; \; 1 \le j \le N \tag{6}$$

$$i = T-1, T-2, \cdots, 1 \quad ; \quad 1 \le i \le N$$

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_{j(y_{t+1})} \beta_{t+1}(j) \tag{7}$$

따라서 F를 최종상태의 집합이라 하고, π_i를 상태 i의 초기확률이라 하면

$$P(Y \mid M) = \sum_{i \in F} \alpha_t(i) = \sum_{i} \beta_t(i) \pi_i \tag{8}$$

이 성립하며, 또한 다음 식도 성립한다.

$$P(y, x_{t-1} = i, x_t = j) = \alpha_t(i) a_{ij} b_{i(y_{t+1})} \beta_{t+1}(j) \quad (9)$$

식(2)에서 식(9)까지의 과정을 Forward-Backward 알고리즘 또는 Baum-Welch 알고리즘이라 한다.[5]

Viterbi Algorithm

- Introduction: $\delta_1(i) = \pi_i b_i(\mathbf{x}_1)$, $\psi_1(i) = 0$
- Recursion: $\delta_{t+1}(j) = \max_{1 \le i \le N} \delta_t(i) a_{ij} b_j(\mathbf{x}_{t+1})$, $\psi_{t+1}(j) = \arg\max_{1 \le i \le N} \delta_t(i) a_{ij}$
- Termination: $P^* = \max_{1 \le i \le N} \delta_T(i)$, $q_T^* = \arg\max_{1 \le i \le N} \delta_T(i)$
- Path backtracking: $q_t^* = \psi_{t+1}(q_{t+1}^*), \quad t = T-1, \ldots, 1$

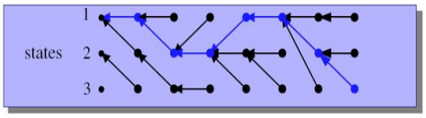


그림 2. Viterbi 알고리즘
Fig. 2. Viterbi Algorithm

두 번째 문제는 관측열 Y가 주어졌을 때 최적의 상태열을 구하는 것이다. 이 문제는 Viterbi 알고리즘을 이용하여 해결할 수 있다.[6]

그림 2에서는 최적의 상태열을 찾는 Viterbi알고리즘의 수식을 나타내고 있다.

2. MAHMM

기존의 모델적응 기법은 입력특징은 수정하지 않고 대신 인식모델을 잡음 상황에 맞도록 적응시키는 방식이다. 현재 거의 모든 음성인식 시스템에서는, Hidden Markov Model(HMM)을 인식모델로 채택하고 있는데 이들 HMM은 많은 양의 오염 되지 않은 음성으로 학습을 한 것이다. 모델적응 기법은 모델적응 기법은 원래, 화자적응을 위하여 쓰이는 기법들로부터 시작되었는데 그 대표적인 예가 maximum a posteriori(MAP) 방법과 maximum likelihood linear regression(MLLR)방법이다. MAP 기법은 적응데이터를 통하여 얻어지는 인식모델과 미리 알고 있는 모델을 interpolation 하는 방식이고, MLLR 기법은 각 인식모델에 적응데이터로부터 구해지는 matrix를 부가하여 변환하는 방식이다.

제안하는 모델적응 방법은 깨끗한 음성과 잡음을 각각 다른 모델(여기서는 HMM)로 표현하고 이 두 모델을 결합하여, 잡음이 섞인 음성의 모델을 생성하는 방법이다.

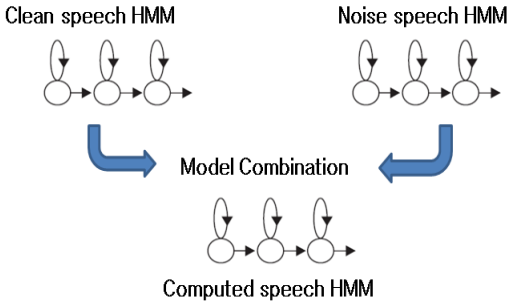


그림 3. MAHMM 알고리즘
Fig. 3. MAHMM Algorithm

본 논문에서는 음성 특징보상 기법 중 Data-driven 기법을 접목한 형태로 조용한 환경에서의 음성데이터와 예측된 잡음모델이 아닌 잡음환경에서의 오염된 데이터를 동시에 사용하여 HMM에서 참조모델을 생성하여 인식하는 방법이다.

3. 실험 및 인식 결과

모바일환경(잡음 60~80dB) 에서의 실험을 위하여 Table 1와 같이 모바일 기기에 사용 가능한 20개의 음절 데이터를 사용 하였고 신호 대 잡음비 각 20dB, 10dB, 5dB 로 실험 하였다.

표 1. 인식단어 리스트
Table 1. Recogntion Word List

NO	음절
1	전화
2	예
3	아니오
4	연결
5	다음
6	이전
7	취소
8	음악재생
9	멈춤
10	정지

11	전곡재생
12	통화
13	검색
14	메뉴
15	영상통화
16	메시지
17	카메라
18	메모
19	사진
20	일정

표 2. HMM 인식률(%)
Table 2. HMM Recognition rate(%)

	참조데이터	S/N Ratio	평균 인식률(%)
HMM	Clean voice data	20 dB	88
		10 dB	48
		5 dB	30

표 3. MAHMM 인식률(%)
Table 3. MAHMM Recognition rate(%)

	참조데이터	S/N Ratio	평균 인식률(%)
MAHMM	Clean voice data + Gaussian noise	20 dB	90
		10 dB	75
		5 dB	62
	Clean voice data + Real life voice data	20 dB	92
		10 dB	84
		5 dB	75

표 2,3에서 알 수 있듯이 기존의 방법에 비해서 약 10%이상 향상된 결과를 보이고 있는데 이는 모바일환경(실제 생활 잡음환경)에서의 경우 Noise데이터를 참조패턴과 접목시켜 인식 주변 환경을 고려한 결과에서이다.

특히 신호 대 잡음비가 10dB이하에서는 인식률의 편차가 큰데 이는 참조패턴이 신호 대 잡음비가 10dB 전후이라고 판단되는 결과로 사료된다.

III. 결 론

본 논문에서 제안한 방법인 MAHMM의 경우 기존의 방법에 비해 모바일환경에서 약 10% 향상된 90%이상의 인식률을 얻을 수 있었다. 특히, 주변 잡음이 심한 자동차 도로 상황에서의 비교 데이터에 대해서도 비교적 양호한 결과를 나타내었는데 이는 주변잡음도 합께 참조데이터에 포함 된 결과이라고 사료되어진다. 그러나, 주변잡음이 상대적으로 많은 지역에서는 실험화자가 다소 크게 발성해야 인식 가능한 레벨에 도달 할 수 있을 것으로 예측되어진다.

참 고 문 헌

[1] 안종영, 김영섭, 김수훈, 허강인, “자동차 ECU제어를 위한 음성인식 패턴매칭 레벨에 관한 연구,” 한국 인터넷 방송통신학회 논문지 10권 1호 pp. 75-80. 2010.
[2] 김남수, 잡음환경에서의 음성인식 Telecommunications Review· 제13권 5호·2003년 10월, p650~661
[3] 中川聖一 ,“確率モデルによる音聲認識”, 電子情報通信學會編, 1989.
[4] 中川聖一, “連續出力分布型HMMによる日本音韻認識”, 音響學會 論文誌 vol. 46, pp.486-496, 1990.
[5] 김수훈, 허강인, “예측차수 변화에 따른 신경망 예측 HMM의 성능비교”, 한국음향학회 학술발표대회 논문집 제16권 1S호, pp.227-231, 1997.
[6] 이종진, “한국어 연속음성 인식시스템의 구현”, 博士學位 論文, 1994.

※ 본 논문은 동아대학교 학술연구비 지원에 의하여 연구되었음.

저자 소개

안 종 영(정회원)

- 1993년 : 동아대학교 전자공학과 공학사
- 1996년 : 동아대학교 전자공학과 공학석사
- 1996-2000 ;현대오토넷 전임연구원
- 2001-2003:한국폴리텍 아산캠퍼스 영상매체과 교수
- 2004-2006 : (주)대성전기 선임연구원
- 현 : 동아대학교 대학원 전자공학과 박사과정, 한국폴리텍Ⅱ 인천대학 컴퓨터정보과 초빙교수

<주관심분야 : 음성신호처리, 임베디드 시스템, DSP, 전장 ECU>

김 영 섭(준회원)

- 2005년 : 동명정보대학교 컴퓨터공학과 공학사
- 2007년 : 동아대학교 전자공학과 공학석사

2009년~현: 동아대학교 전자공학과 박사과정

<관심분야 : 패턴인식, 음성/영상처리, DSP application>

김 수 훈

- 1990년: 동아대학교 전자공학과 공학사
- 1992년 : 동아대학교 전자공학과 공학석사
- 1999년 : 동아대학교 전자공학과 공학박사
- 2001년~현: 부천대학 모바일통신과 부교수

<주관심분야 : DSP, 음성인식, 모바일콘텐츠>

허 강 인 (교신저자)

- 1980년 : 동아대학교 전자공학과 공학사
- 1982년 동아대학교 전자공학과 공학석사
- 1990년 경희대학교 전자공학과 공학박사
- 1998년 9월~1989년 8월 일본 쓰쿠바대학 객원연구원
- 1992년 9월~1993년 8월 일본 도요하시대학 객원연구원
- 1984년-현: 동아대학교 전자공학과 교수

<주관심분야 : DSP, 음성인식, 음성합성, 신경회로망>