고립단어 음성인식에서 신경망을 이용한 사용자 적응형 후처리

김영진°, 김은주**, 김명원**
숭실대학교 정보과학대학 컴퓨터학과
liebulia@yahoo.co.kr°, blue7788@naver.com, mkkim@computing.ssu.ac.kr

User Adjustment Post-Process Using Neural Network In Isolated Word Speech Recognition

Youngjin Kim° Eunju Kim** Myoungwon Kim**
School of Computing, Soongsil University

요 약

최근 PDA나 PMP와 같은 개인용 모바일 기기의 인터페이스 개발로써 잡음환경에 강인한 음성인식 기술들이 연구되고 있으며 이러한 방법으로 오류패턴, 순차패턴, 의미정보, 문맥정보와 같이 인식기에 독립적인 정보를 이용하거나 영상 정보와 같이 언어와 성격이 다른 이질적인 정보를 이용하여 후처리를 하는 연구들이 진행되어 왔다. 그러나 인식기와 독립적인 정보로 후처리를 하는 방법들의 인식률은 인식기의 사전 인식률이 주변 잡음에 의해 떨어질 경우 후처리 인식률도 같이 떨어지는 현상이 벌어진다. 따라서 본 논문에서는 주변 잡음으로 인한 인식기의 사전 인식률에 저하를 줄이는 방법으로 사용자 적응형 후처리를 제안한다. 사용자 적응형 후처리에 사용되는 데이터는 사용자의 발화에 대한 인식기의 출력 값들이며, 출력 값들은 화자독립모델에 의해 계산되는 각 단어들의 유사도 들이다. 따라서 화자독립모델의 결과를 사용자 적응형 후처리에 적용한 결과 인식기의 오류를 58.7% 줄일 수 있었다.

1. 서 론

최근 모바일 시장의 급성장으로 인하여 PDA나 PMP와 같은 모바일 기기에 알맞은 다양한 인터페이스들이 개발·상용화되고 있다. 음성인식 또한 이러한 인터페이스 중 하나로 상용화하기 위하여 많은 연구들이 진행 되고 있으며, 이미 상용화 된 제품들도 있다. 하지만 모바일 기기의 사용은 장소의 구애를 받지 않아 다양한 잡음이 마이크로 유입되게 되며 이러한 잡음으로 인하여 많은 오인식이 유발된다. 이러한 현상은 음성인식을 모바일 기기의 인터페이스로 활용하는데 많은 어려움을 준다.

현재 음성인식은 다화자의 음성을 인식하게 하기위하여 HMM[1,2]으로 화자독립모델을 만들어 인식기에 사용한다. 하지만 화자독립모델은 다화자의 음성데이터를 수집해야 하기 때문에 시간과 비용이 많이 발생하게 되고, 화자의 상태에 따른 변형된 발음이나 특정 지역성 발음까지 포함하기는 더더욱 힘든 일이다. 따라서 이를 보완하기 위하여 오류패턴비교(error pattern matching)[3]나 순차패턴(sequence pattern)[4], 문맥정보(context information)[5,6]와 같은 후처리를 적용하여 인식결과를 보정한다. 그러나 이러한 후처리 방법들은 사전 음성인식 확률의 신뢰도가 높다는 가정 하에 진행되므로 사전 음성인식의 상태에 많은 영향을 받게 된다. 즉, 학습되어진 화자독립모델의 인식률이 낮을 경우 후처리 적용이 어렵다.

본 논문에서는 화자독립모델의 인식률을 높이기 위한 후처리 방법으로 특정 사용자에 적응해가는 방법을 제안한다. 이는 개인 휴대용으로 발전해 가는 모바일 기기의 특성을 이용한 것이며, 개인의 특성이 많이 반영될수록 좋은 인식률을 보이기 때문이다. 그리고 사전에 만들어진 화자독립모델의 특성도 동시에 고려되어야 한다. 이는 화자독립모델 생성 시 사용되어진 발화데이터에 따라 모델 내의 노드들에 확률 값이 달라지고 이로 인한 최종 인식 단어에 미치는 영향이 달라지기 때문이다. 따라서 사용자의 특성과 화자독립모델의 특성을 동시에 고려하기위한 속성으로 사용되어지는 것이 각 단어들의 우도(likelihood) 값들이다. 그리고 우도 값들의 패턴학습을 위하여 신경망을 사용하였다.

* 본 연구는 산업자원부에서 지원하는 "수퍼지능칩 및 응용기술 개발"과제의 지원에 의해 수행되었습니다.

2. 관련연구

2.1. 독립적인 정보를 이용한 후처리

주변 잡음에 강인한 음성인식을 위하여 연구되고 있는 음성인식 후처리 기법은 다음과 같은 것들이 있다. 먼저, 오류패턴비교(error pattern matching)[3]는 자주 발생이 되는 오류를 패턴으로 저장하고 이를 인식에 적용하는 방법이다. 따라서 음성인식기가 최종적인 인식 결과를 내기 전에 오류패턴 정보로 가중치를 보정 하거나 인식 가능한 단어 리스트에 추가하는 방법으로 사용된다. 두 번째로, 순차패턴(sequence pattern)[4]은 단어들 간의 순서를 미리 정하고 이를 이용하여 인식 결과에 가중치를 부여하는 방식이다. 여기서 순차패턴은 문법과 같은 단어의 배열 순서를 정하거나 특정 도메인에서 사용되는 경로들을 표현하여 사용하게 된다. 마지막으로, 문맥정보(context information)[5,6]는 문맥상의 내용을 기반으로 사용 단어들을 구축하고 이를 이용하여 인식 결과에 가중치를 부여하는 방식이다. 즉 뜻이나 쓰임새가 유사한 단어들 끼리 묶어 가중치에 적용한다.

2.2. 신경망을 이용한 문맥정보 후처리

[7]에서 신경망을 이용하여 문맥의 패턴정보를 학습시키는 후처리 방법을 제안하였다. 이 방법은 발화 데이터와 독립적인 문맥정보를 이용함으로 최종 인식의 보정 비율을 높일 수 있으며, 패턴학습에 강한 신경망학습을 사용함으로서 발화 데이터

의 잡음 현상에 강인한 인식 결과를 가져올 수 있다. 하지만, 이 방법 역시 이전 단어의 인식에 영향력이 크게 작용하게 되어, 표준모델의 인식률이 현저하게 떨어질 경우 보정한 결과의 신뢰도가 떨어지게 된다.

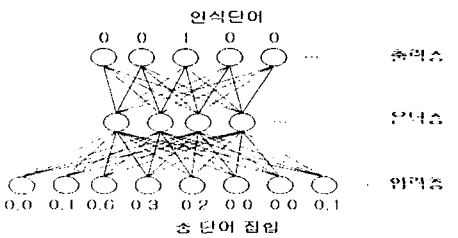
3. 신경망을 이용한 사용자 적응형 후처리 모델

본 장에서는 한정된 명령어 음성인식에 있어 기본이 되는 화자독립모델의 인식률을 높이기 위한 후처리 방법과 그에 따른 사용자 적응형 후처리 모델의 생성 기법에 대해 기술한다.

3.1. 사용자 적응형 후처리 모델(APNN)

일반적으로 발화자는 발화자만의 특별한 억양을 가지게 되며 이러한 현상은 지방 사투리 억양의 경우 더더욱 화자독립모델과의 차이를 나타낸다. 따라서 특별한 억양을 가지는 음성이 입력으로 들어올 경우 화자독립모델에서는 오인식을 유발하게 된다. 이는 고립단어 인식에 있어 입력으로 들어오는 음성을 기기에서 사용되는 단어들에 근사시키는 방식을 사용하고 있기 때문이다. 따라서 자주 오류가 발생되는 단어에 대해서는 좀더 세부적인 검토가 필요하다. 본 논문에서는 화자독립모델과 적응형 모델을 분리하여, 두 번의 공정과정을 거쳐 최종 인식률을 올리는 방법을 제안한다. 이는 억양의 성격이 다르거나 주변 잡음이 많은 데이터들을 화자독립모델에 학습 시킬 경우 일반적인 인식률을 저하시키거나 APNN의 의존도가 너무 높아지는 현상이 벌어지기 때문이다. 또한, 모바일 기기의 기능 또는 명령어의 추가가 발생할 경우 APNN을 재학습 시키는 것을 막기 위해서 이다.

APNN의 생성은, 화자독립모델의 인식 결과와 사용자의 피드백을 가지고 각 단어의 오류가 얼마나 발생하는지를 체크하여 일정 수치 이상이면 모델 단어로 선정하고, 선정된 단어로 후처리 모델 학습을 진행한다. 학습에 사용되는 입력 데이터는 화자독립모델의 결과 값인 각 단어들의 우도(likelihood)값이고 이 우도 값들의 패턴을 학습시키기 위해 <그림 1>과 같은 신경망을 구축하였다.



<그림 1> 사용자 적응형 후처리 모델(APNN)

APNN은 입력, 은닉, 출력층으로 구성된 신경망의 전형적인 형태를 가진다. 입력층은 모든 단어의 수와 일치하며, 각각의 노드에 입력되는 값은 화자독립모델에서 출력되는 각 단어들의 우도 값들을 정규화한 값들이 된다. 그리고 출력층의 노드는 화자독립모델에서 자주 오류를 범하는 단어와 그 대상이 되는 단어들의 셋으로 구성된다. 그리고 인식 시 모델이 선택 되었을 경우 출력층의 결과를 최종 결과로 하게 된다.

3.2. 특징 추출

화자독립모델의 인식은 HMM 모델을 기반으로 각 단어의 우도들을 출력하도록 할 수 있다. 이러한 우도들은 입력되는 발화자의 음성이 어떠한 단어들과 얼마나 유사한 가를 나타내는 측도가 되며, 이를 유사도라 한다. 따라서 음성인식 전처리 인식기는 유사도의 값이 가장 큰 값을 선택하게 되지만, 실생활 속의 잡음이 섞이거나 지역적 발음이 입력으로 들어올 경우 유사도의 값들이 틀어지는 현상이 자주 발생하게 된다. 그러나 이러한 틀어짐 현상 속에서도 모든 단어들의 유사도 들을 보면 일정 패턴이 존재함을 실험결과 알게 되었다. 즉, 발화자가 의도한 단어가 적은 유사도 값을 나타내거나 전혀 감지를 못한다 하더라도 주변 단어에 이치는 유사도를 보면, 이 음이 무슨 음인지 알 수 있게 된다. 따라서 발화 음성의 각 유사도 값들을 특징 값으로 추출하여 APNN의 학습 데이터로 선정 한다.

3.3. 정규화

각 단어의 유사도들을 일정한 공간에 두고 비교하기 위하여 정규화를 수행한다. <식 1>은 사용자 후처리 모델의 입력 값을 일정 공간에 근사시키는 최대-최소 정규화 식이다.

$$\frac{(Node-Min)}{|Max-Min|}|\max-\min|+\min \qquad \text{<식 1>}$$

학습 데이터에서 가장 큰 값을 Max로 하고, 가장 작은 값을 Min으로 하였다. 이는 각각의 인식 단어들이 주변 환경의 음에 서로 영향을 받기 때문이다. $\max$는 허용 범위의 최대이고, $\min$은 최소이며 이들은 각각 1.0, 0.0 이 된다.

3.4. 모델 생성 기준

APNN은 오류를 자주 일으키는 단어를 찾고, 오인식에 자주 등장하는 단어들을 찾아 그 단어들 간의 패턴으로 재분류하는 방법을 사용한다. 따라서 <식 2>와 같은 방식으로 모델 단어를 선정하게 된다.

$$\frac{n(R)}{n(T)} > 1 \qquad \text{<식 2>}$$

<식 2>에서 $n(T)$은 해당 단어의 총 발화 횟수이고, $n(R)$은 인식된 횟수이다. 이때, 인식의 유무는 따지지 않는다. 이 방식으로 선정된 단어에는 목적이 다른 단어들을 가지게 되어 출력 노드에는 모델 단어와 목적이 다른 단어들로 구성된다. 단, 노드 결정은 <식 3>을 만족해야 한다.

$$\frac{n(N)}{n(M)} > 0.1 \qquad \text{<식3>}$$

<식 3>에서 $n(M)$은 모델 노드의 수이며, $n(N)$은 목적이 다른 단어의 수이다. 이는 아주 드문 환경적 요인에서 발생할 수 있는 경우를 배제하기 위한 임계값이다.

4. 실험 및 평가

4.1. 실험환경 설정 및 실험 방법

본 실험에서 화자독립모델은 음성인식에서 일반적으로 많이 사용되는 HMM 모델을 사용하며, 이는 CMU에서 HMM 모델 학습용으로 만든 HTK(HMM tool kit)을 사용하여 생성한다. 또한 화자독립모델의 학습에 사용되어진 음성 데이터는 남녀노소 약 100명의 화자로부터 모은 데이터를 사용하였고 실험에 선정된 단어들은 일반적인 상황을 고려하여 특정 명령어 형태가 아닌, 일반적으로 사용되어지는 단어 75개를 선정 하여 <표 1>과 같이 구축 하였다.

실질적으로 사용자는 초기 모든 단어에 대해 같은 비율로 발화하는 행위를 하도록 한다면 사용자는 불편함을 크게 느끼게 된다. 그러나 본 실험에서는 어떠한 단어에서 사용자의 적응형 모델이 생성될지 예상할 수 없기 때문에 모든 단어에 대해 실험하는 것으로 하였다. 따라서 초기에는 75개의 단어를 10번씩 발화하여 각 단어들의 인식률을 체크한 뒤 초기 모델을 생성한다. 생성은 오류가 많이 발생했던 단어들만을 선정하게 되며, 선정된 단어는 발화자가 10번씩 다시 발화한다. 이러한 학습을 반복하면 점점 사용자에게 맞는 적응형 모델이 생성되게 된다. 그리고 매 실험은 APNN을 만들고 평가하는 것임으로, 실시간으로 발화하고 인식률을 체크하는 방식으로 진행 하였다. 그리하여 <표 2>와 같은 모델들은 얻을 수 있었으며, 이

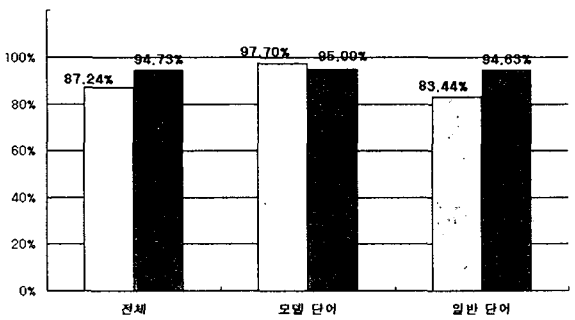
데이터는 총 20번의 반복학습을 통하여 얻은 결과이다.

<표 1> 75개의 선정 단어

No	단어	No	단어	No	단어	No	단어	No	단어
1	가구	16	글	31	된장	46	아들	61	웬일
2	가보	17	꿀	32	들깨	47	애기	62	육성
3	가족	18	나	33	등쌀	48	약속	63	의사
4	간식	19	날뛰다	34	딸	49	양	64	자리
5	간판	20	남산	35	마음	50	역사	65	잣새
6	갈치	21	납기	36	멥새	51	연못	66	잣송이
7	감기	22	농비	37	목	52	예	67	줄기
8	감자	23	늑대	38	바퀴	53	옷	68	찌개
9	값이	24	다리	39	밥	54	옷밥	69	창
10	고삐	25	달	40	보리	55	완수	70	칼
11	곡식	26	돌다리	41	비행	56	왔다	71	투구
12	괜찮다	27	동백	42	뿔	57	왼쪽	72	풀
13	구리	28	동이	43	사람	58	욕	73	하나
14	구었다	29	동쪽	44	셈	59	용산	74	획기적
15	귤	30	동태	45	쌀	60	원고	75	훔러

<표 2> 생성된 사용자 적응형 후처리 모델(APNN)

모델명	노 드 명
가구	가구,가보,간판,하나,
가보	가보,감자,연못,
감기	감기,남산,
구리	구리,뿔,
글	글,사람,
납기	날뛰다,납기,바퀴,
늑대	괜찮다,날뛰다,늑대,동태,들깨,등쌀,멥새,왔다,
달	달,딸,사람,쌀,
동태	남산,농비,동백,동태,등쌀,용산,
목	나,마음,목,연못,욕,육성,투구,
보리	보리,뿔,
아들	아들,풀,
애기	간식,고삐,날뛰다,농비,다리,돌다리,동이,된장, 들깨,바퀴,비행,애기,예,웬일,줄기,
약속	약속,역사,육성,
예	고삐,늑대,딸,비행,사람,쌀,양,예,
옷	밥,옷,육성,
옷밥	구었다,옷밥,왔다,
완수	마음,완수,의사,잣송이,
찌개	값이,찌개,획기적,
투구	괜찮다,농비,돌다리,동이,된장,등쌀,마음,비행, 용산,원고,투구,하나,



<그림 2> 화자독립모델과 사용자 적응형 후처리 모델의결과 비교

<그림 2>는 화자독립모델의 인식률과 자주 발생이 되는 오인식 단어의 유사도 패턴을 이용하여 APNN을 생성한 이 모델을 적용시킨 결과 <그림 2>와 같은 결과를 얻었다.

<그림 2>는 화자독립모델과 APNN의 결과를 비교한 그래프이다. '전체' 항목은 전체 인식률을 비교한 것이고, '모델 단어' 항목은 모델로 선정된 20개의 단어를 비교한 것이다. 그리고 '일반 단어' 항목은 모델 단어로 선정되지 않은 단어들을 비교한 것이다. 여기서 '모델 단어' 항목은 APNN의 인식률이 떨어지는 현상이 벌어지는데, 이는 모델로 선정된 단어의 성격으로 벌어지는 현상이다. 즉, 모델로 선정된 단어는 인식률이 다른 단어들에 비해 인식률이 높은 단어들로 선택 되며, 본 단어의 인식률과 다른 단어가 본 단어로 오인식 된 비율를 포함하여 선정된 단어이다. 따라서 본 논문은 오인식으로 인하여 모델 단어들로 잘못 인식 되어진 단어들을 원래의 단어로 인식 되도록 보정하는 것이 된다. 그리고 이러한 인식률은 반복적인 APNN의 학습을 통하여 높아지는 것을 확인하였으며, 그 결과 약 58.7%의 오류 수정률을 보였다.

5. 결론 및 향후 연구

실험 결과, 잡음 환경에서 화자독립모델에 APNN을 적용하여, 최종 인식률을 높이는 효과를 가져왔으며, 계속적인 반복학습을 통하여 인식률이 높아지고 있음을 확인하였다. 하지만, 학습 방법에 있어 사용자의 학습 발화 데이터를 다시 요구해야 하는 상황이 생기게 되며, 이러한 시간과 노력은 최종사용자에게 불필요한 것으로 인식 될 수 있다.

따라서, 적응형 모델을 학습시키는데 필요한 학습데이터를 사용 중에 실시간으로 얻을 수 있는 방법들이 연구 되어져야 하며, 학습의 반복 횟수를 줄일 수 있는 방법 또한 같이 연구 되어져야 한다.

참 고 문 헌

[1]Deller, Hansen, Proakis, "Discrete-Time Processing Of Speech Signals", IEEE PRESS, pp677~744, 2000.

[2]M. Ostendorf, "From HMM's to segment models: a unified view of stochastic modeling for speech recog- nition", IEEE SPA. pp.360-378, 1996.

[3]Satoshi Kaki, Eiichiro Sumita, and Hitoshi Iida, "A Method for Correcting Speech Recognition Using the Statistical features of Character Co-occurrence.", COLING-ACL, pp.653-657, 1998

[4]J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal and MC. Hsu., "PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth." ICDE, pp.215-224, 2001

[5]Minwoo Jeong, Byeongchang Kim, Lee, G.G., "Semantic-oriented error correction for spoken query processing", ASRU IEEE Workshop on, pp.156-161, 2003

[6]Myung Won Kim, Joung Woo Ryu, Eun Ju Kim, "Speech Recognition by Integrating Audio, Visual and Contextual *feature Based on Neural Networks*", *International* Conference on Natural Computation, LNCS 3614, pp.155 ~ 164, 2005

[7]Szu Chen Jou, Tanja SChultz, Alex Walbel, "WHISPERY SPEECH RECOGNITION USING ADAPTED ARTICULATORY REATURES", IEEE International Conference on Volume 1, March 18-23, pp.1009-1012, 2005.

[8]송원문, 김은주, 김명원, "사용자 발화 순차패턴을 이용한 음성인식 후처리", 한국정보과학회 한국컴퓨터종합학술대회 2005, Volume.32No.01 pp.0709~0711, 2005.7.

[9]권오욱, 박준, 황규웅, "의사 형태소 단위 대어휘 연속음성인식기 개발", 제 15회 음성통신 및 신호처리 워크샵 논문집, pp.320-323, 1998