문맥정보를 이용한 이중모드 음성인식

류정우[0] 김은주 김명원
숭실대학교 컴퓨터학부
ryu0914@bilab.ssu.ac.kr[O], mkim@comp.soongsil.ac.kr

Bimodal Speech Recognition Using Contextual Feature

Joung Woo Ryu,[0] Eun Ju Kim, Myung Won Kim
School of Computing, Soongsil University

요 약

최근 잡음환경에서 신뢰도 높은 음성인식을 위해 음성정보와 영상정보를 융합하는 이중모드 음성인식 방법이 활발히 연구되고 있다. 본 논문에서는 보다 음성 인식률을 향상시키기 위해 사용자가 말한 단어들의 순차 패턴을 나타내는 문맥정보를 이용한 후처리 방법을 제안한다. 이러한 문맥정보를 인식하기 위해 다층퍼셉트론 구조를 갖는 문맥정보 인식기를 제안한다. 이중모드 음성인식기와 문맥정보 인식기 결과를 효율적으로 결합하기 위한 후처리 방법으로 순차 결합방법을 제안한다. 문맥정보를 이용한 이중모드 음성인식이 잡음 환경에서 90%이상의 인식률을 보였다. 본 논문은 잡음환경에서 강인한 음성인식을 위해 문맥정보와 같은 사용자 행동패턴이 새로운 정보로 이용될 수 있다는 가능성을 제시한다.

1. 서 론

최근 들어 사회가 점차 멀티미디어화 됨에 따라 인간과 기계의 인터페이스를 좀 더 간편하고 명확하게 실현하기 위하여 얼굴표정이나 방향, 입술모양, 응시추적, 손동작 그리고 음성 등을 이용한 다중모드(multi-modal)형태의 인식 연구가 활발히 진행되고 있다.

특히, 이러한 연구는 최근 이동 단말기의 기술이 발전함에 따라 잡음환경에 강인한 음성인식 방법인 이중모드(bimodal) 음성인식 방법으로 활발히 연구되고 있다. 이중모드 음성인식 방법이란 잡음환경에 민감한 음성정보를 보완할 수 있는 영상정보를 동시에 고려함으로써 음성인식률을 향상시키는 방법이다. 예를 들어, 공장과 같은 시끄러운 환경에서 대화할 때 사람들은 서로의 음성뿐만 아니라 입모양 혹은 제스쳐와 같은 영상정보를 이용하여 음성을 인식하는 경우를 생각할 수 있다.

이러한 이중모드 인식기로는 HMM(Hidden Markov Model)과 신경망이 일반적으로 많이 사용된다. [2]에서는 HMM을 이용하여 인식하기 전에 음성과 영상 특징을 융합하였다. 이와 같이 인식하기 전에 융합하는 방법을 특징 융합방법이라고 하며 이것은 음성과 영상정보의 표본비율(sampling rate)이 다르기 때문에 융합하기가 어렵다. 이러한 동기화 문제를 해결하기 위해 저주파 통과 보간법(low-pass interpolation)을 사용하여 표본을 추출하였고, 새로운 특징은 10msec가 중복된 25msec윈도우로부터 생성하였다. 그러나 HMM을 이용한 융합 방법에 있어 결과에 민감한 반응을 주는 학습 변수인 상태(state) 수와 가우시안 혼합(Gaussian mixture) 수를 결정하기 어렵고, 특히 일반적으로 사용되는 CDMM(Continuous Density Hidden Markov Model)은 입력특징들이 확률적 독립성 조건을 만족해야 하는 제약사항들이 있어 적용하기 어렵다[1][2].

신경망 중 TDNN(Time-Delay Neural Network)은 음소의 지속시간 및 음성 신호 내의 시제 위치 등 다양한 조건에서도 상당히 정확하게 음소를 인식할 수 있는 신경망 모델이다[5]. MS-TDNN(Multi State TDNN)은 DTW(Dynamic Time Warpping)층을 추가하여 연속 단어를 인식할 수 있도록 TDNN을 확장한 모델이다[6][7]. 이러한 MS-TDNN을 이용하여 음성정보와 영상정보를 융합한 이중모드 MS-TDNN이 개발되었다[3].

이중모드 MS-TDNN은 두 단계 학습과정을 통해 모델이 형성된다. 첫 번째 학습과정은 음소단위로 이루어지며 음성정보와 영상정보 각각에 대해 독립적인 TDNN 인식기를 생성한다. 두 번째 학습과정은 고립단어 단위로 DTW에서 가장 적합한 단어에서부터 각 TDNN 출력층까지 역전파(backpropagation) 알고리즘을 통해 학습이 이루어진다.

이와 같이 이중모드 MS-TDNN은 음소레벨에서 단어를 인식해야 하므로 시간 축 변화(time axis variation) 문제를 해결하기 위한 DTW 알고리즘이 요구된다. 그러므로 보다 복잡한 모델이 생성될 뿐만 아니라 잡음에 민감하고 음소간의 구분이 어렵다는 음소인식의 문제점을 그대로 가지게 된다.

또한, [4]에서는 잡음환경에서 숫자음을 인식하기 위해 음성과 영상정보를 융합한 이중모드 인식기를 적용하고 있다. 적용된 이중모드 인식기는 두 정보를 특징 융합방법으로 융합하기 위해 단지 입력층에 영상정보를 위한 노드를 추가한 다층퍼셉트론으로 설계되었다. 따라서 모델에 대한 견고성(robustness)은 좋으나 모델 크기가 커짐에 따라 계산량이 증가되는 비효율적인 문제점을 가진다.

[5]에서는 이러한 문제점을 보완하기 위해 이질적인 정보들을 효율적으로 융합할 수 있는 신경망을 이용하고, 효율적으로 모델을 생성할 수 있는 고립단어 인식 모델인 이중모드 신경망(BMNN : BiModal Neural Network)을 제안하였다. BMNN은 4개층으로 이루어진 다층퍼셉트론의 구조를 가지며 각 층은 입력특징의 추상화 기능을 수행한다. 특히 세 번째 층인 융합층은 잡음에 의한 음성 정보의 손실을 보상하기 위하여 음성과 영상특징을 통합하는 기능을 수행한다.

본 논문에서는 잡음환경에서 음성인식률을 보다 향상시키기 위해 사용자가 말한 단어들의 순차 패턴을 나타내는 문맥정보를 이용한 후처리 방법을 제안한다. 이 때 [5]에서 제안하고 있는 이중모드 음성인식기인 BMNN을 인식기로 사용한다.

본 논문의 구성은 다음과 같다. 2절에서는 잡음환경에서 음성인식률을 향상시키기 위해 제안한 문맥정보를 이용한 후처리 방법에 대해 서술한다. 3절에서는 본 논문에서 제안한 방법에 대한 실험 및 분석 결과를 서술하며, 4절에서는 결론 및 향후 연구에 대해 기술한다.

2. 문맥정보를 이용한 후처리 기술

2.1 문맥정보 인식기

사용자 명령어 사용패턴과 같은 순차 패턴을 인식하기 위해 <그림 1>과 같은 문맥정보 인식기(context recognition)를 제안한다. 일반 신경망은 입력노드에 순차 정보를 가지고 있지 않으므로 순차 패턴을 인식할 수 없다. 따라서 제안된 문맥정보 인식기는 다층퍼셉트론으로 설계하였으며 입력 층에 순차적 정보를 주기 위해 입력 값을 이진형(0과 1)으로 표현하였다.

예를 들어 인식할 명령어(단어)가 "영", "일", "이", "삼" 네 개라고 가정할 경우, 문맥정보 인식기는 <그림 1>과 같이 설계될 수 있다. 여기서 입력노드 개수는 인식할 명령어(단어) 개수에 선행 단어 개수를 곱한 것과 같고 출력노드 개수는 인식할 명령어(단어) 개수와 같게 설정한다. 선행단어 개수란 현재 단어를 예측하기 위해 고려되는 단어의 개수로써, 앞서 인식한 명령어(단어)들 중 순차적으로 가장 최근에 인식한 단어의 개수를 말한다. 만약 입력값이 "일영이 : 0100 1000 0010" 으로 입력되면 출력노드에는 "일영이" 다음에 가장 많이 사용되는 해당 명령어(단어) 노드에 가장 높은 값이 출력된다. <그림 1>에서는 "삼" 이 가장 높기 때문에 "일영이" 다음에 사용될 명령어(단어)를 "삼"으로 예측하게 된다. 따라서 선행단어 개수를 설정할 때, 너무 크게 설정하게 되면 패턴에 따라 발생빈도가 낮기 때문에 예측 값이 낮은 경향을 보이는 반면, 너무 작게 설정하게 되면 예측 값이 극소의 패턴으로 편중되는 경향을 보이게 된다.

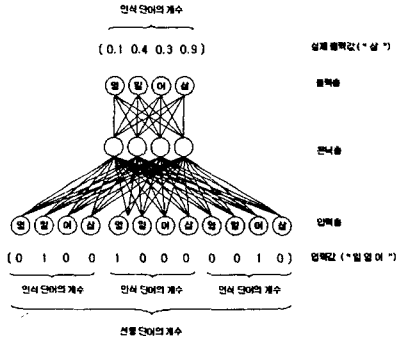


그림 1 문맥 정보 인식기

3.2 문맥정보를 이용한 후처리

잡음 환경에서 보다 강인한 음성인식기를 개발하기 위해 본 논문에서는 문맥 정보를 이용한 후처리 방법을 제안한다. 후처리 방법이 적용된 음성인식기 구조는 <그림 2>와 같다. 본 논문에서 사용한 이중모드 음성인식기는 [5]에서 제안한 BMNN 인식기를 사용하였다. 따라서 최종 인식 결과는 BMNN인식기의 출력 값과 문맥정보 인식기의 출력 값을 결합함으로써 나타나게 된다. 그러므로 BMNN 인식기와 문맥정보 인식기는 독립적으로 학습을 수행하여 모델을 생성하였다.

두 인식기 결과를 효율적으로 결합하기 위한 방법으로 <그림 3>와 같은 순차 결합(sequential combination) 방법을 제안한다. 제안한 결합방법은 음성인식기의 인식결과가 사용자가 설정한 임계값(θ) 보다 작을 경우 문맥정보 인식 결과를 고려하는 방법이다. 만약 두 인식기의 결과가 모두 임계값(θ)보다 작을 경우 입력 정보에 대해 정확한 인식이 이루어지지 않았다고 보고 두 인식기의 출력 값을 곱함으로써 출력 값의 차이가 적은 것을 선택할 수 있도록 한다. 각 인식기의 출력 값은 [0,1]의 값을 갖으며, 1에 가까울수록 출력 값의 신뢰성이 높다는 것을 의미한다. 임계값(θ)은 사용자에 의해 결정되며 사용자가 인식기의 결과를 신뢰할 수 있는 최소 한계 값을 의미한다.

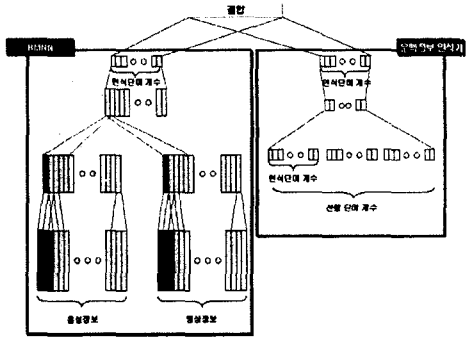


그림 2 문맥 정보를 이용한 음성 인식기

$BMNN(O_i)$: BMNN에서 i번째 출력노드의 출력 값
$Con(O'_i)$: 문맥정보 인식기에서 i번째 출력노드의 출력 값
θ: 임계값

if ($\theta < BMNN(O_i)$)
  $i = \max_i(BMNN(O_i))$ 번째 고립단어로 인식
else if ($\theta < Con(O'_i)$)
  $i = \max_i(Con(O'_i))$ 번째 고립단어로 인식
else if (($BMNN(O_i) \le \theta$) and ($Con(O'_i) \le \theta$))
  $i = \max_i(BMNN(O_i) \cdot Con(O'_i))$ 번째 고립단어로 인식

그림 3 순차 결합 방법

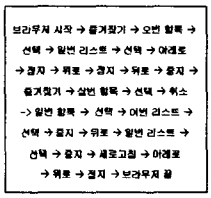
4. 실험 및 결과

실험에 사용된 데이터는 35개의 고립단어를 27회 발음한 화자 종속 데이터이다. 단어들은 이동 단말기에서 사용될 수 있는 명령어들로 구성되었다. 잡음 환경에서의 음성신호를 생성시키기 위해 가우시안 잡음(20db, 10db, 5db)을 인위적으로 추가하여 잡음 데이터를 생성하였다.

본 논문에서 사용하고 있는 음성 특징 추출방법과 영상 특징 추출 방법은 기존의 방법들로서 음성 특징 추출 방법에 ZCPA(Zero Crossing with Peak Amplitude)[6] 방법을, 영상 특징 추출 방법에는 PCA(Principle Component Analysis) 방법을 사용하였다.

본 실험에서 사용될 BMNN의 모델 구조는 고립단어 인식을 위해 입력 프레임을 64프레임(프레임당 10ms)으로 설정하였고 각 프레임에서 16차원의 특징을 추출하였다. 입력층의 윈도우 크기는 음소를 표현하기에 충분한 30ms인 3프레임으로 설정하였고 중복 영역 크기는 2프레임으로 설정하였다. 은닉층의 윈도우 크기는 더 넓은 시제 영역을 학습할 수 있도록 5프레임으로 설정하였고 중복 영역 크기는 4프레임으로 설정하였다. 따라서 은닉층은 62프레임으로 응합층은 58프레임으로 설정하였다.

후처리 방법에 사용할 문맥정보 인식기를 생성하기 전에, 이동 단말기 상에서 사용자가 <그림 4>와 같이 사용하는 순차적 명령어 패턴이 존재한다고 가정한다. 이때 <그림 4>-(a)와 같은 순차적 명령어 패턴이 사용자가 사용하는 전체 명령어 패턴들 중에서 발생되는 비율이 70%, 50%, 30%등으로 다르게 하여 각각의 학습데이터들을 생성하였다. 예를 들어 <그림 4>-(a)에서처럼 "브라우저 시작, 즐겨찾기, 오번항목" 이라고 명령한 다음 "선택" 이라는 명령어를 제시할 경우를 7회 발생시키고 3회는 선행단어들을 임의로 선택하여 생성함으로써 70%의 규칙성을 갖는 학습데이터를 생성하였다. 이와 같이 비율을 다르게 하여 학습데이터를 생성한 이유는 문맥정보 인식기가 학습데이터에 포함하고 있는 특정 패턴의 비율에 민감하게 반응하는지 알아보기 위해서이다.



(a) 사용자가 사용하는 명령어 패턴

NO.	입력값 (선행단어)			출력값
1	-	-	-	브라우저 시작
2	-	-	브라우저 시작	즐겨찾기
3	-	브라우저 시작	즐겨찾기	오번 함목
4	브라우저 시작	즐겨찾기	오번 함목	선택
••	•••			•••
30	아래로	위로	정지	브라우저 끝

(b) 테스트 데이터

그림 4 사용자가 사용하는 명령어 패턴과 테스트 데이터

본 실험에서 사용할 문맥정보 인식기는 명령어(단어)를 예측하기 위한 선행단어 개수를 3개로 설정하였다. 따라서 각 층의 노드의 개수는 입력층 105개, 은닉층 52개, 출력층 35개로 설정하였다. 입력층은 인식할 단어 개수가 35개이고 선행단어 개수가 3개 이므로 105개의 노드로 설정되었다. 반면, 은닉층의 노드 개수는 실험을 통해 얻어졌다.

생성된 문맥정보 인식기의 성능을 확인하기 위한 테스트 데이터는 <그림 4>-(a)패턴을 사용하여 생성하였다. 테스트 데이터는 <그림 4>-(b)와 같이 사용자가 처음 사용하는 명령어 "브라우저 시작" 부터 마지막으로 사용하는 명령어 "브라우저 끝" 까지 총 30개의 데이터로 구성되었다. 이렇게 생성된 테스트 데이터로 실험한 결과 70%, 50%, 30% 모델에 대한 인식률이 83.33%, 83.33%, 86.67%임을 확인할 수 있었다. 여기서 "브라우저시작", "즐겨찾기", "오번 항목" 같이 선행단어가 전부 존재하지 않는 단어들에 대해서는 모든 모델들이 올바르게 인식하지 못하였다. 또한 앞에서 기술한 바와 같이 학습데이터를 생성할 때 임의적으로 생성된 패턴에 따라 성능의 차이가 다소 발생하였다. 그 이유는 임의적으로 발생한 패턴들 때문에 선행단어는 같으나 예측할 단어가 틀린 경우가 발생함으로써 올바른 학습이 이루어지지 않았기 때문이다. 그러나 비율에 상관없이 비슷한 성능을 보이고 있음을 확인할 수 있다. 따라서 문맥정보 인식기는 일정 이상의 비율을 갖는 패턴들을 학습한다는 것을 알 수 있다.

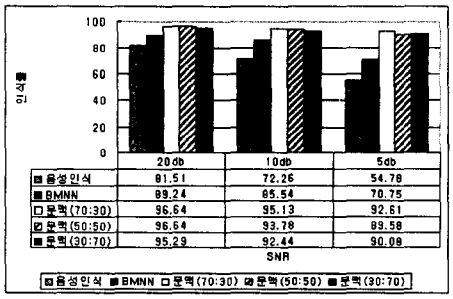


그림 5 문맥정보를 이용한 실험 결과

이와 같이 학습된 문맥정보 인식기를 <그림 2>와 같이 후처리 방법에 적용하였다. 문맥정보 인식기는 특정 사용자의 명령어 사용패턴을 학습하기 때문에 화자 종속에만 적용된다. <그림 5>은 실험결과를 보여준다. 결과를 살펴보면 잡음환경에서 음성 평균인식률은 69.51%이고 BMNN의 평균인식률은 81.84%인 반면, 문맥 정보를 이용한 후처리 방법을 적용하였을 경우 평균 인식률은 93.57%로 가장 높은 결과를 보인다. 또한 잡음 증가에 따른 인식에 대한 평균 감소율을 살펴보면 음성인식 경우 13.36% 평균 감소율을 보이고 BMNN 경우 9.24% 평균 감소율을 보이고 있으나, 문맥 정보를 이용한 후처리 방법을 적용할 경우 2.72% 평균 감소율을 보임으로써 보다 잡음에 영향을 받지 않는다는 것을 확인할 수 있다. 이처럼 만약 사용자의 순차 명령 패턴이 존재한다면 이러한 정보들을 문맥정보 인식기로 학습함으로써 잡음환경에서 보다 우수한 성능을 보일 수 있다는 가능성을 제시한다. 순차 명령 패턴이 존재한다는 가정 하에서 제안한 방법에 의해 성능이 향상된 이유는 음성과 영상정보만으로 구별될 수 없었던 패턴들을 구별할 수 있었기 때문이다. 그러나 순차 명령 패턴이 존재하지 않는다면 문맥정보가 성능을 향상시켜준다고 보장 할 수는 없다. 따라서 본 논문에서는 잡음환경에서 명확한 음성인식을 위해 문맥정보와 같은 사용자 행동패턴이 새로운 정보로 이용될 수 있다는 가능성을 제시한다.

5. 결론 및 향후 연구

본 논문에서는 잡음 환경에서 음성 인식률을 향상시키기 위해 사용자가 사용하는 명령어들의 순차 패턴을 나타내는 문맥정보를 이용한 후처리 방법을 제안한다. 문맥정보를 이용한 이중모드 음성인식이 잡음 환경에서 90%이상의 인식률을 달성하였다. 따라서 제안한 문맥정보와 같은 사용자 행동패턴이 잡음환경에서 강인한 음성인식을 위해 고려할 수 있는 새로운 정보로 이용될 수 있다는 가능성을 제시한다.

향후 연구로는 본 논문에서 제시한 문맥정보를 이용한 후처리 방법을 화자 독립에서도 적용될 수 있도록 문맥정보 표현에 대한 연구를 진행할 것이다. 또한 음성인식기와 문맥정보 인식기 결과를 결합하는데 있어, 본 논문에서는 순차 결합 방법을 제안하여 음성인식기 결과를 문맥정보 인식기 결과보다 항상 더 중요하게 반영함으로써 상황에 따른 적응성이 떨어질 수 있다. 따라서 상황에 따라 두 결과에 대한 판단 가중치를 고려할 수 있도록 신경망, HMM 혹은 퍼지이론을 적용한 보다 일반적인 결합방법에 대해 연구를 진행할 것이다.

6. 참고 문헌

[1] Gemello, R.; Albesano, D.; Mana, F.; Moisa, L.; " Multi-source neural networks for speech recognition: a review of recent results" , Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on , vol. 5, pp. 265-270, 2000.

[2] Xiaozheng Zhang; Merserratt, R.M.; Clements, M.; , " Bimodal fusion in audio-visual speech recognition " , Image Processing. 2002. Proceedings. 2002 International Conference on ,vol.1, pp. 964-967, 2002..

[3] C.Bregler, S.Manke, H.Hild and A. Waibel, "Bimodal sensor integration on the example of "speech-reading", Proc. of IEEE Int. Conf. on Neural Networks, San Francisco, 1993.

[4] 이상원, 박인정, "잡음환경에서 음성-영상 정보의 통합 처리를 사용한 숫자음 인식에 관한 연구", 전자공학회논문지, 제38권 CI편, 제3호, pp.61-67, 2001년 5월.

[4] 류정우, 성지애, 이순신, 김명원, "신경망을 이용한 이중모달 음성 인식 모델링", 한국정보과학회 춘계학술발표논문집(B), 제30권, 제1호, 2003년 4월.

[6] Doh-Suk Kim,Soo-Young Lee, Rhee M. Kil, "Auditory Processing of Speech Signals for Robust Speech Recognition in Real-World Noisy Environments", IEEE Trans. on Speech and Audio Processing, vol.7, no.1, pp. 55-69, January 1999.