김일성종합대학학보
(자연과학)
주체105(2016)년 제62권 제8호

JOURNAL OF **KIM IL SUNG** UNIVERSITY
(NATURAL SCIENCE)
Vol. 62 No. 8 JUCHE105(2016).

음소음성인식에서 심층신뢰망을 리용한 한가지 음향모형화방법

리정철, 현성군

위대한 령도자 **김정일**동지께서는 다음과 같이 교시하시였다.

《정보산업을 빨리 발전시키고 인민경제의 모든 부문을 정보화하여야 합니다.》

(《**김정일**선집》 증보판 제20권 380페지)

심층신뢰망(DBN: Deep Belief Network)은 수많은 신경세포(일명 세포)와 중간층(일명 은폐층)을 가지고있는 심층신경망의 일종으로서 맨 웃 2개층사이에 무방향련결을 가지고 나머지층사이에는 방향성련결을 가지는 일종의 그라프모형이다.

최신자동음성인식체계는 전형적으로 HMM(Hidden Markov Model)을 리용하여 음성신호의 련속구조를 모형화하고있으며 국부적인 스펙트르변동은 가우스밀도혼합을 리용하여 모형화하고있다. 그러나 HMM은 가우스혼합모형의 제한된 표현능력과 비현실적인 조건부독립성가정으로 하여 음성신호안에서 호상작용하는 지식원들을 잘 모형화하지 못한다. 지난 몇십년동안 HMM의 이러한 제한성을 극복하기 위한 많은 방법들이 제안되여 일정한 성과를 거두었지만 완전한 성공은 기대하지 못하고있다. 반면에 신경망은 구조가 깊어질수록 비선형모형화능력이 강해지는것으로 하여 우점을 가지지만 은폐층수가 많아질 때 학습효률이 떨어지고 속도가 매우 느린것으로 하여 현실에서 많이 리용되지 못하고있다.

최근 심층신경망의 파라메터들을 학습시키기 위한 효률적인 알고리듬들과 하드웨어기술의 급속한 발전으로 하여 필기체문자인식, 3-D대상인식, 정보검색, 운동포착자료모형화 등 패턴인식분야에서 그 적용효과가 성공적으로 나타나기 시작했으며 부분적으로 음성인식에 적용하기 위한 연구들도 진행되고있다.

이러한 심층신경망의 일종인 심층신뢰망의 우점은 학습자료의 통계적구조를 각이한 준위에서 표현할수 있는 강력한 능력을 가지고있다는것이다.

론문에서는 심층신뢰망[1]의 강력한 학습자료표현능력을 음성인식에서의 음성스펙트르변동모형화에 리용하기 위한 한가지 방법을 제안하였다.

1. 심층신뢰망을 리용한 음향모형화

1) 심층신뢰망의 입력층

론문에서 리용한 심층신뢰망은 제한된 볼츠만기계를 구축블로크로 하는 다층신경망이다. 이러한 심층신뢰망으로 음성의 스펙트르변동을 모형화하기 위하여서는 이미 보편적으로 리용되고있는 MFCC에 기초한 파라메터나 LPC에 기초한 파라메터 혹은 그것의 변종들을 망의 입력신호로 리용할수 있는데 그 입력신호는 다음과 같이 정규화되여야 한다.

$$X_t' = \frac{X_t - m}{\sigma} \tag{1}$$

여기서

$$m = \frac{\sum_{t=1}^{T} X_t}{T}, \quad \sigma = \sqrt{\frac{\sum_{t=1}^{T}(X_t - m)^2}{T}}$$

이고 T는 발성문장에서 추출된 특징파라메터수이다.

MFCC에 기초한 39차원특징파라메터를 리용하는 경우 입력신호 X_t는 령평균 및 단위분산을 가지도록 정규화되여야 한다.

또한 특징파라메터들의 련관을 모형화하면서 고정된 입력차원을 가진 심층신뢰망을 음소인식에 적용하기 위하여 특징벡토르들중 n개의 련이은 프레임들로 구성된 문맥창문으로 심층신뢰망의 맨 아래층(입력층 혹은 로출층)에 있는 세포의 상태들을 설정한다. 이때 입력층세포의 개수는 총 39×n이다.

2) 심층신뢰망의 은폐층

심층신뢰망에서 은폐층의 은폐세포들은 원래의 입력자료의 련관성특징들을 각이한 준위에서 포착하고 표현하게 된다.

제한된 볼츠만기계들을 탄창식으로 쌓아 심층신뢰망을 구축하는 방법의 기본원리는 볼츠만기계에 의하여 학습된 모형파라메터 θ가 $p(v|h,\ \theta)$와 사전분포 $p(h|\theta)$를 둘다 정의하므로 θ를 학습한 후에 $p(v|h,\ \theta)$는 유지되게 하면서 $p(h|\theta)$를 더 좋은 모형으로 교체시키자는것이다.

이를 위해 우선 다른 볼츠만기계를 맨 웃층에 삽입한 다음 그 아래층에서 출력된 숨은 활성벡토르들을 삽입된 볼츠만기계를 훈련시키기 위한 입력자료로 리용한다. 다음 삽입된 은폐층의 무게들을 초기화하기 위하여 점차적인 층별CD(Contrast Divergence)법[2]을 리용한다. 이 과정을 n번 반복하면 n개의 은폐층이 점차적으로 구축되며 초기화된 파라메터들은 판별학습을 통하여 다시 세밀조절된다.

이런 방식으로 구축된 심층신뢰망은 DBN-DNN(Deep Neural Network)이라고도 부른다.

3) 심층신뢰망의 출력층

론문의 목적이 음소인식을 진행하자는데 있으므로 출력층에 음소들을 반영하는 변수들을 추가할수 있다. HMM의 견지에서 구체적으로 보면 음소들은 어떤 정상구간에서의 스펙트르변동을 모형화하는 상태들의 시간적이행으로 묘사되므로 매 세포가 그 상태를 서술하도록 출력층을 설계할수 있다. 이로부터 매 음소에 3개의 상태를 할당하는 경우 출력층의 세포수를 음소수×3으로 설정할수 있다. 그런데 출력층에 반영할 가능한 상태들이 많고 매 상태들의 빈도분포가 균등하지 못할 때 망의 출력층에서 최대로 활성화된 1개 상태만을 취하는 표준SOFTMAX부호화[2]보다 출력층의 서로 다른 몇개의 부호들을 리용하여 사후확률을 계산하고 이에 기초하여 상태를 결정하는것이 더 우월할수 있다.

이를 위해 이제 모형의 출력세포가 q개 있다고 할 때 q차원 부호벡토르 z를 가지고 매 상태를 표현한다고 하자. 그리고 C_j가 상태 j의 부호를 취하는 행벡토르라고 하면 z가 주어질 때 모형이 상태 s에 할당하는 사후확률은 다음과 같이 된다.

$$p(s\,|\,\theta,\ \ z)=\frac{e^{C_s z}}{\sum_j e^{C_j z}} \tag{2}$$

이 방법은 매 상태당 1개의 부호비트를 리용하는 표준SOFTMAX법(SSM)의 일반화로 볼수 있으므로 우리는 일반화된 SOFTMAX법(GSM)으로 부르기로 한다.

2. 실험 및 결과분석

1) 실험조건

음소인식실험은 음성인식성능검사를 위한 표준음성코퍼스(TIMIT)[3]에 대해서 진행되였다. 훈련모임에 포함된 발성자수는 462명이며 모형파라메터조절을 위해 리용된 개발모임에는 50명의 발성자들을 포함시켰다. 최종인식결과를 얻기 위하여 24명의 검사모임을 리용하였다. 이때 우리는 음성자료를 10ms의 프레임속도를 가진 25ms하밍창문을 리용하여 해석하였다. 모든 실험들에서는 12차MFCC와 에네르기, 그것들의 1계 및 2계미분을 함께 리용하여 음성을 표현하였으며 자료는 령평균 및 단위분산을 가지도록 정규화하였다. 론문에서는 망의 입력신호를 구성하기 위해 11개 프레임으로 이루어진 문맥창문을 리용하고 망의 출력층에는 183개(61개 음소×3개 상태)의 클라스(매 클라스는 1개의 상태를 의미)기호를 리용하였다. 그러므로 망출력세포들을 183개 목표클라스들에 대한 확률로 변환시키게 된다.

복호화후에 시작무음과 마감무음은 제거되였고 61개의 음소클라스들은 득점평가를 위하여 39개의 클라스로 변환되였다.

한편 모든 실험들에서는 훈련모임으로부터 추정된 음소2그람언어모형을 리용하였다.

다른 한편 복호기파라메터들은 살창탐색의 매 과정에 개발모임의 성능을 최량화하기 위하여 조절되였다.

2) 평가실험

먼저 모형에 층당 2 048개의 은페세포를 배치하고 은페층수의 변화에 따르는 영향을 조사하였는데 그 결과는 그림 1과 같다.

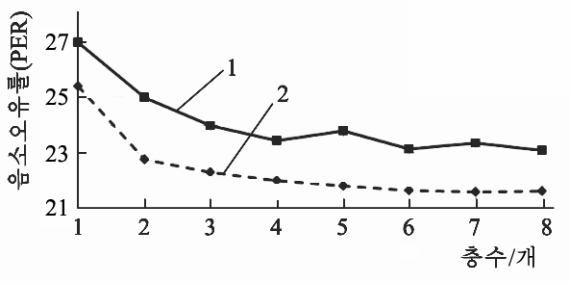

그림 1. 모형깊이가 음소오유률에 주는 영향
1－검사모임일 때, 2－개발모임일 때

그림 1에서 보여준것처럼 두번째 은페층을 추가하였을 때 음소오유률이 크게 축소되며 4이상부터 음소오유률에서의 변화정도는 별로 차이가 없다는것을 알수 있다.

따라서 은페층수를 4로 고정시키고 다른 계산지표들에 대한 실험을 진행하였다.

이 모형에서 최적인 출력층의 세포수 즉 GSM법에 의한 사후확률계산에 리용되는 부호벡토르차원은 128정도였다.

다음으로 은페층크기(세포수)에 따르는 음소오유률을 조사하였다.(표 1) 표 1에서 보다싶이 성능이 층당 세포수에 따라 크게 달라지지 않기때문에 은페층당 2 048개의 세포를 리용하여 실험을 진행하고 그에 대한 계산지표들에 관심을 두었다.

망의 출력층에 리용된 GSM법의 영향을 검사하기 위하여 그 성능을 SSM법과 비교하였다. 실험결과 128차원GSM과 183차원SSM을 리용하는 2개의 구조모형은 개발모임에 대

해서 모두 음소오유률이 22%였지만 검사모임에 대해서는 각각 23.36, 23.90%였다.

음소오유률개선의 기본리유를 명백히 하기 위하여 128차원GSM모형의 마지막층에 각각 128개, 2 048개의 세포를 가진 은폐층을 추가한 5층 심층신뢰망들과 비교하였다.(표 2)

표 1. 층크기가 음소오유률에 주는 영향

세포수	개발모임일 때/%	검사모임일 때/%
1 024	21.94	23.46
2 048	22.00	23.36
3 072	21.74	23.54

표 2. GSM이 음소오유률에 미치는 영향

모형	개발모임일 때/%	검사모임일 때/%
128차원 GSM	22.00	23.36
128개 은폐세포	22.00	23.00
2 048개 은폐세포	21.98	23.73

표 2에서 보는바와 같이 128개 은폐세포를 리용하였을 때가 성능이 제일 높은데 이것은 심층신뢰망이 마지막층의 좁은 통로가 서로 다른 클라스들간에 특징들을 공유하도록 해주면서 과정합을 막을수 있게 한다는것을 보여준다. 다시말하여 출력층무게들을 제외한 나머지무게들이 모두 예비훈련되므로 학습된 좁은 출력층은 예비훈련의 효과를 얻지 못하는 파라메터수를 실질적으로 감소시킨다고 말할수 있다.

다음으로 최적인 마지막 좁은 층의 세포수를 결정하기 위한 실험결과를 그림 2에 주었다.

그림 2의 개발모임그라프에서 보는바와 같이 5번째 은폐층에 적어도 64개의 은폐세포를 배치할 때까지는 오유률에서 별로 차이가 없다는것을 알수 있다. 또한 검사모임그라프에서 나타난 많은 차이는 통계적으로는 의미가 없다해도 5번째 층의 크기를 4번째 층에 비하여 조금 축소시키는것이 좋다는것을 암시한다.

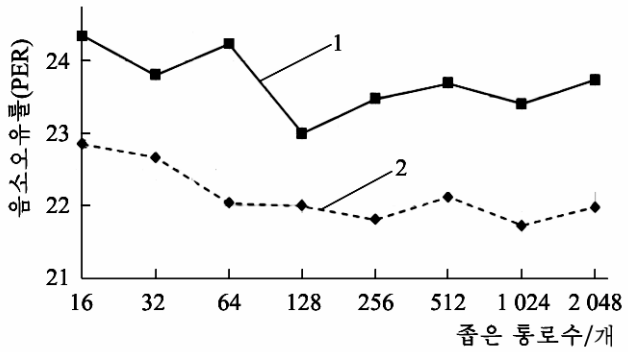

그림 2. 좁은 통로크기가 음소오유률에 주는 영향
1－검사모임일 때, 2－개발모임일 때

맺 는 말

심층신뢰망을 음소음성인식을 위한 음향모형화에 적용하고 이 모형들의 구조상 깊이와 은폐층크기가 음소오유률에 미치는 영향을 조사하였다. 그리고 실험을 통하여 심층신뢰망에서 마지막 좁은 층이 과정합을 막는데 좋다는것을 확증하였다.

참 고 문 헌

[1] G. E. Hinton et al.; Neural Computation, 18, 1527, 2006.

[2] G. E. Hinton et al.; IEEE Signal Processing Magazine, 11, 2, 2012.

[3] G. D. Thomas et al.; Journal of Artificial Intelligence Research, 12, 263, 1995.

주체105(2016)년 4월 5일 원고접수

An Acoustic Modeling Method based on Deep Belief Networks in the Phone Speech Recognition

Ri Jong Chol, Hyon Song Gun

We proposed a method to apply DBNs(Deep Belief Networks) to acoustic modeling for speech recognition, which these are recently proved to be very effective for a variety of machine learning problems.

Key words: speech recognition, restricted boltzmann machine, deep belief networks