김일성종합대학학보
정보과학
주체109(2020)년
제66권 제1호

조선어음성합성체계실현을 위한 심층신경망의 입력층파라메터구성방법

리세웅, 한철진

지난 시기 음성합성분야에서 단위음결합방식과 HMM(Hidden Markov Model)에 기초한 통계적파라메터에 의한 음성합성방식을 많이 리용하였다. 특히 HMM에 기초한 음성합성방식은 체계의 용량이 작고 합성음의 특징을 단순한 파라메터조작으로 쉽게 변경시킬수 있다는 우점을 가지고있다. 그러나 이 방법에서는 모형파라메터들이 지나치게 평활화되는것과 같은 결함을 비롯하여 여러가지 원인으로 합성음질이 떨어지는 결함이 있다.

최근 인공지능분야에서 심층학습기술을 많이 리용하고있으며 음성처리분야에서도 심층학습기술을 받아들이기 위한 연구[1, 2]가 활발히 진행되고있다.

론문에서는 심층신경망을 리용한 조선어음성합성체계실현에서 신경망의 입력층파라메터를 구성하기 위한 한가지 방법을 제안하였다.

1. 심층신경망을 리용한 음성합성체계의 구성

음향모형(Acoustic Model)을 심층신경망으로 구성한 음성합성체계를 그림 1에 보여주었다.

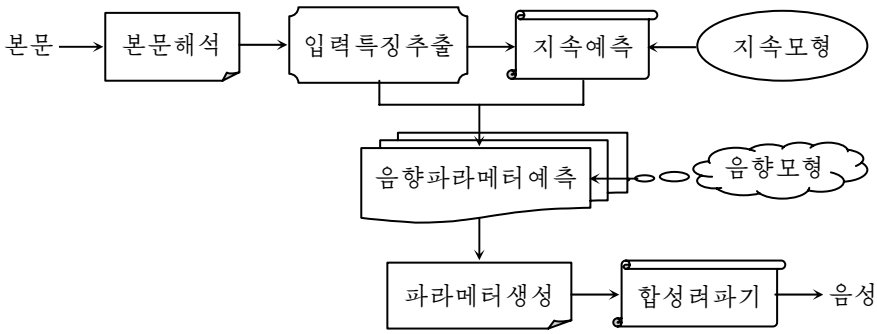

그림 1. 음향모형을 심층신경망으로 구성한 음성합성체계

그림 1에서 보여준것처럼 지속모형과 음향모형은 심층신경망모형이며 각각 본문특징으로부터 개별적인 음소들의 지속길이와 음향파라메터들을 예측하게 된다.

신경망의 구조를 그림 2에 보여주었다. 그림 2에서 x_i는 신경망의 입력층유니트이고 h_i^j는 j번째 숨은층의 i번째 유니트이며 y_j는 신경망의 출력층유니트를 의미한다. 매 유니트는 여러개의 입력을 받아 1개의 출력을 진행하며 출력 z는 다음과 같이 계산된다.

$$u = \sum_{i=1}^{I} w_i x_i + b, \quad z = f(u) \tag{1}$$

여기서 w_i는 유니트의 무게, x_i는 유니트의 입력, b는 유니트의 편위, $f(\cdot)$는 유니트의

활성화함수이다.

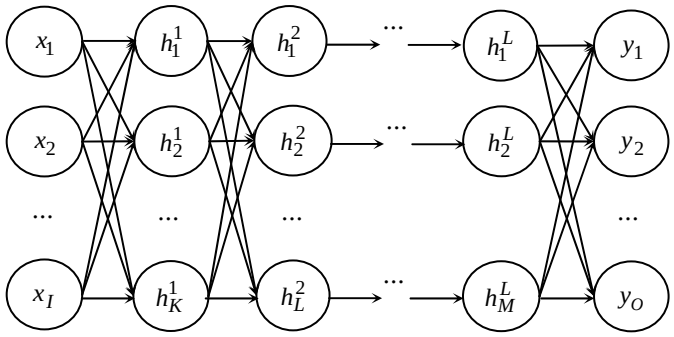

그림 2. 신경망의 구조

신경망모형의 입력층과 출력층파라메터를 어떻게 구성하는가에 따라 음성합성프로그람의 자연성이 좌우되게 된다. 지속모형과 음향모형의 출력층파라메터는 각각 음소의 지속길이와 음성부호화기의 결과로 얻어지는 음향파라메터를 리용하고 입력층파라메터는 본문특징을 리용한다. 이로부터 론문에서는 조선어음성합성을 위하여 리용되는 본문특징들과 그것을 리용하여 입력층파라메터를 구성하기 위한 한가지 방법을 제안하고 실험을 진행하여 그 효과성을 검증하기로 한다.

2. 입력층파라메터구성

음성합성을 위한 음향모형과 지속모형의 입력은 본문특징이며 심층신경망으로 모형화하는 경우 입력을 구성하는 방법에 대하여 론의하기로 한다.

음성의 합성단위를 음소로 정하고 음소의 음향학적특징에 영향을 줄수 있는 문맥인자들을 설정하였다.

문맥인자를 표 1에 보여주었다.

표 1. 문맥인자

기 호	설 명
$p1$	현재음소의 전전음소
$p2$	현재음소의 전음소
$p3$	현재음소
$p4$	현재음소의 다음음소
$p5$	현재음소의 다음다음의 음소
$p6$	음절내에서의 음소위치(앞방향)
$p7$	음절내에서의 음소위치(뒤방향)
$s1$	단어내에서 음절위치(앞방향)
$s2$	단어내에서 음절위치(뒤방향)
$s3$	음절의 모음음소이름
$s4$	음절에 강조억양이 있는가?(1/0)
$s5$	이전음절의 모음음소이름
$s6$	이전음절에 강조억양이 있는가?(1/0)

기 호	설 명
*s*7	다음음절의 모음음소이름
*s*8	다음음절에 강조억양이 있는가?(1/0)
*w*1	단어내의 음절개수
*w*2	억양구내에서 단어위치(앞방향)
*w*3	억양구내에서 단어위치(뒤방향)
*w*4	단어의 마지막형태부품사정보
*w*5	단어의 마지막토정보
*w*6	단어뒤에 붙은 문장기호
*h*1	억양구내의 음절개수
*h*2	억양구내의 단어개수
*h*3	문장에서 억양구위치(앞방향)
*h*4	문장에서 억양구위치(뒤방향)
*h*5	억양구의 끝음조
*e*1	문장의 음절개수
*e*2	문장의 단어개수
*e*3	문장의 억양구개수

여기서 *p*1－*p*5까지는 음소이름으로서 ㄱ, ㄴ, ㅏ, ㅑ 등을 가질수 있으며 그 가지수는 음소개수와 같다. 또한 *s*4는 0 혹은 1의 값을 가질수 있으며 *e*1~*e*3과 같이 개수정보를 나타내는것으로 하여 임의의 값을 가질수 있는 특징도 있다. 즉 문맥인자들은 그 값범위특징에 따라 세가지 부류로 가를수 있다. 하나는 *p*1－*p*5와 같이 제한된 특징값을 가지는것이고 다른 하나는 *s*4와 같이 2진값을 가지는것이다. 그리고 *e*1과 같이 임의의 수값을 가지는것으로 갈라볼수 있다. 두번째와 세번째 인자들인 경우 그 특징값을 그대로 입력파라메터로 리용하면 되지만 첫번째 경우는 이 특징값을 수값형태로 변화시켜 입력하여야 한다. 그것을 위해 매 특징형태에 1, 2, … 순서로 번호를 붙여 1차원의 수값으로 리용할수도 있고 HMM에 기초한 음성합성에서 리용하던 질문들을 리용하여 다차원의 2진값으로 리용할수도 있다.

실례로 음소이름인 경우 첫번째 방법인 경우 ㄱ에는 1, ㄴ에는 2, ㄷ에는 3 등을 대입시켜 리용하며 두번째 방법인 경우 100개의 질문에 대한 대답으로 얻어지는 100차원의 2진벡토르를 리용하게 된다.

리용될수 있는 질문형태들은 다음과 같다.

현재음소가 모음인가?
현재음소가 자음인가?
음소가 홑모음인가?
현재음소가 앞모음인가?
현재음소가 뒤모음인가?
현재음소가 긴모음인가?
현재음소가 열린모음인가?
현재음소가 코소리인가?
현재음소가 터침소리인가?
…

추가적으로 음향모형의 경우 음소지속을 나타내는 정보가 더 포함된다. 학습단계에서는 이 정보로 HMM을 리용한 상태준위정렬결과를 리용하며 합성단계에서는 그림 1에서 보여준것처럼 지속모형의 출력(예측된 지속정보)을 리용한다. 이러한 방법으로 추출된 입력층파라메터는 0~1사이값을 가지도록 정규화된다.

정규화는 학습자료의 최소, 최대값에 기초하여 다음의 식을 리용하여 진행한다.

$$\hat{x} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{2}$$

실천에서는 보다 좋은 학습을 위하여 정규화된 값의 범위를 [0.01, 0.99]로 제한한다. 즉 정규화식은 다음과 같다.

$$\hat{x} = (0.99 - 0.01)\frac{x - x_{\min}}{x_{\max} - x_{\min}} + 0.01 \tag{3}$$

3. 실 험 결 과

론문에서는 16kHz, 16bit, Mono방식의 파형자료(830개 문장, 약 3h분)를 학습자료로 리용하였으며 음성부호기는 WORLD[3]를 리용하였다. 심층신경망실험을 위하여 학습자료를 학습, 확인, 검사모임의 세가지 부분으로 분할하였다. 학습모임은 730문장, 확인 및 검사모임은 각각 50문장으로 구성하였다.

론문에서는 음향모형과 지속모형의 두가지 모형의 숨은층을 4개 층으로 구성하였으며 매 층의 유니트수는 512개로 설정하였다. 또한 학습률은 0.01, 훈련반복수는 25로 설정하고 모형훈련을 진행하였다.

실험에서는 HMM에 기초한 음성합성체계와 심층신경망을 리용한 음성합성체계를 같은 학습자료를 리용하여 구축하고 합성음질의 비교평가를 진행하였다. 또한 심층신경망을 리용한 음성합성체계는 두가지로 구성하였다. 하나는 입력층파라메터구성에서 음소이름이나 품사와 같은 부류특징들을 순서화된 1차원수값으로 대응시켜 구성한 경우(DNN 1)이고 다른 하나는 질문에 대한 결과로 얻어지는 다차원2진벡토르로 구성한 경우(DNN 2)이다.

합성음질평가는 입력문장 10개를 주고 5명의 전문가가 합성음들에 0~5점사이의 점수를 주고 평균점수를 계산하여 평가하는 방식으로 진행하였다.

심층신경망을 리용한 음성합성체계의 성능평가를 표 2에 보여주었다.

표 2. 심층신경망을 리용한 음성합성체계의 성능평가

체계	점수 1	점수 2	점수 3	점수 4	점수 5	총평
선행방법(HMM)	3.8	3.85	3.9	3.9	3.8	3.85
제안한 방법(DNN 1)	3.9	4	3.9	3.9	3.9	3.92
제안한 방법(DNN 2)	4.1	4	4	4.1	4	4.04

표 2에서 보여준것처럼 심층신경망을 리용한 음성합성체계의 성능이 HMM에 기초한 음성합성체계보다 더 높다는것을 알수 있다. 또한 입력특징량구성에서 DNN 1의 경우가 DNN 2의 경우에 비해 합성음질이 더 낮다. 그것은 DNN 1의 경우 부류별특징을 순서화된 1차원수값으로 대응시키면 서로 류사한 특징값들을 반영할수 없기때문이다.

맺 는 말

심층신경망을 리용한 조선어음성합성체계실현에서 신경망의 입력층파라메터를 구성하기 위한 방법을 제안하고 그 성능을 실험을 통하여 검증하였다.

참 고 문 헌

[1] William Chan et al.; ICASSP, 4960, 2016.

[2] H. Zen et al.; ICASSP, 7962, 2013.

[3] M. Morise et al.; IEICE Transactions on Information and Systems, 99, 7, 1877, 2016.

주체108(2019)년 11월 5일 원고접수

A Method of DNN's Input Layer Parameterization for the Korean Text-to-Speech Synthesis System

Ri Se Ung, Han Chol Jin

In this paper, we present a neural network input layer parameterization method to implement Korean text-to-speech synthesis system based on a DNN, and verified the effectiveness by experiment.

Keywords: DNN, speech synthesis, statistical speech synthesis