ISSN 2005-8063
2016. 9. 30.
Vol.8 No.3
pp. 31-38

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

http://dx.doi.org/10.13064/KSSS.2016.8.3.031

심층신경망을 이용한 조음 예측 모형 개발*

Development of articulatory estimation model using deep neural network

유 희 조 · 양 형 원 · 강 재 구 · 조 영 선 · 황 성 하 · 홍 연 정 · 조 예 진 · 김 서 현 · 남 호 성**

You, Heejo · Yang, Hyungwon · Kang, Jaekoo · Cho, Youngsun · Hwang, Sung Hah · Hong, Yeonjung · Cho, Yejin · Kim, Seohyun · Nam, Hosung

Abstract

Speech inversion (acoustic-to-articulatory mapping) is not a trivial problem, despite the importance, due to the highly non-linear and non-unique nature. This study aimed to investigate the performance of Deep Neural Network (DNN) compared to that of traditional Artificial Neural Network (ANN) to address the problem. The Wisconsin X-ray Microbeam Database was employed and the acoustic signal and articulatory pellet information were the input and output in the models. Results showed that the performance of ANN deteriorated as the number of hidden layers increased. In contrast, DNN showed lower and more stable RMS even up to 10 deep hidden layers, suggesting that DNN is capable of learning acoustic-articulatory inversion mapping more efficiently than ANN.

Keywords: the Wisconsin X-ray Microbeam Database, speech inversion, artificial neural network, deep neural network

1. 서론

말소리, 즉 음성(speech)은 혀를 비롯한 여러 조음 기관(articulators)의 체계적이고 물리적인 움직임, 즉 '조음'(articulation)을 통해 만들어진다. 최근 공학, 교육, 의료 등 다양한 분야에서 조음 연구의 성과가 활용됨에 따라, 조음 연구의 중요성과 조음 정보의 유용성은 더욱 강조되고 있다.

이와 같은 조음 정보에 대한 연구의 한 축이 음향과 조음 간 매핑(acoustic-to-articulatory mapping) 내지는 speech inversion에 대한 분야이다. Speech inversion은 말소리의 음향 정보(acoustic information)를 바탕으로 그 소리를 생성해 낸 조음 형태(articulatory configuration)를 역으로 재구성하는 기술로, 최근 음성과 관련된 다양한 연구 분야에서 활용되고 있다. 특히 자동 음성 인식(automatic speech recognition 이하 ASR) 분야와 음성 합성(speech synthesis)분야에서는 널리 적용되고 있으며, 이 외에도 아직은 미비하지만 제2언어 학습자를 위한 효율적인 발음 교육, 의료적 차원으로는 청력이나 뇌 기능, 그리고 발음 기관상의 문제를 겪는 환자들의 조음 치료에도 유용하게 이용될 수 있다 [1][2][3][4].

이처럼 speech inversion은 다양한 학문에 접목하여 활용할 수 있는 잠재력을 가지지만 [5]의 연구에서도 언급하다시피 speech inversion이 본질적으로 갖고 있는 문제점으로 인해 쉽게 기술 발전을 이루지 못하였고 따라서 그 뛰어난 활용성에도 불구하고 오랜 기간 정체되었다. 다음은 본 논문에서 언급하고자 하는 speech inversion의 세 가지 문제점이다.

첫째, 하나의 음향 매개변수 셋은 특정 조음 형태에만 대응되

* 이 논문은 2015년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2015S1A5A8017748)
** 고려대학교, hnam@korea.ac.kr, 교신저자
Received 30 May 2016; Revised 11 July 2016; Accepted 20 September 2016

지 않고, 다양한 조음 형태에 대응될 수 있다(one-to-many problem). 예를 들어, 인간의 발화 중에는 휴지 구간(pause)이 존재할 수 있는데, 이 휴지 구간에도 조음기관들은 계속적으로 다양한 변이형을 형성할 수 있다. 이는 무음이라는 음향 파라미터 셋에 다양한 조음 형태가 대응될 수 있음을 보여주고 있다. 또 다른 예로, 'perfect memory'란 단어를 발화할 때 /t/라는 음소는, 음향적으로 실현되지는 않지만, 실제 조음상에서는 /t/ 음소를 발화하기 위한 혀의 움직임이 관찰 된다.

둘째, 음향과 조음 사이에는 비선형(non-linear)의 관계가 존재한다. 발화 중에 일어나는 조음동작들은 분절 할 수 없는 연속적인 움직임들의 시간적 중첩(temporal overlap)으로 구성된다. 하지만 음향에서는 조음의 연속성이 선형적으로 드러나지 않는다. 예를 들어, /s/ 마찰음을 만들어내기 위해서는 혀의 끝을 치경(alveolar)쪽에 최대한 가까이 위치시키되 약간의 틈을 남겨두어야 한다. 혀는 /s/를 발음하기 위해 해당 조음 위치로 서서히 이동하게 되지만, 그 약간의 틈이 형성되기까지 /s/ 음소는 발화되지 않는다. 이처럼 /s/는 연속적으로 음성을 형성해 나가지 않고 혀와 치경사이가 특정한 틈을 이루는 구간에서 급격하게 실현된다. 이러한 비선형성(non-linearity)은 음향과 조음의 매핑을 더욱 어렵게 만든다. 이러한 비선형상의 관계는 혀의 위치와 공명주파수 사이에서도 나타난다.

셋째, 동시조음(coarticulation)도 음향과 조음의 복합적인 관계를 보여준다. 즉, 조음은 연속성을 갖고 있기 때문에 조음동작들 간의 중복이 필연적으로 발생하게 된다. 예를 들어, eighth에서 /t/는 치경을 건드리며 나는 음소임에도 불구하고 뒤따르는 치음 /ɵ/에 동화되어 치아 부근에서 발화 된다. 이처럼 인간의 발화 속에는 무수히 많은 음소들이 인접한 음소들의 영향을 받게 되고, 이러한 관계 속에서 동시조음은 필연적으로 발생하며 이를 음향상에서 확인하는 것은 매우 힘들다.

이와 같은 문제에 대해 제시된 한 가지 해결책으로 기계학습의 일종인 인공 신경망(artificial neural network 이하 ANN)이 도입되었다. 인공 신경망 모형은 데이터를 기반으로 모델을 수립하며, 일대다 대응(one-to-many mapping), 즉 비고유하고(non-unique) 비선형적인(non-linear) 변수들 간의 관계를 적절히 포착하는 특징이 있어, speech inversion의 난제 해결의 실마리를 제공 한다 [6].

[7]은 speech inversion을 기계학습을 이용해 구현한 초기의 연구로서, ANN의 원시적 모델인 multi-layer perceptron(MLP)를 사용하여 X-ray Microbeam Database(XRMB)데이터로부터 영어 파열음 6개의 조음 움직임을 예측해냈다. 구체적으로 CV음절로 된 반복된 녹음 데이터를 사용하여 ANN 모델을 구축하였으며, 94~98%의 인식 정확도를 보였다. [8], [9]에서는 ANN을 이용하여 음성으로부터 조음기관 위치를 예측하려는 시도가 이루어졌으며, [10]에서는 다양한 타입의 /r/ (bunched 및 retroflex)을 대상으로 conditional density modes 방식을 사용하여 조음기관을 예측 하려고 하였다. [11]에서는 deep belief network(이하 DBN)를 사용하여 speech inversion문제를 해결하려고 시도하였으며 실제 조음 정보가 담긴 'mngu0' 조음 코퍼스를 이용하여 음향과 조음관계를 매핑하였다. 결과로 얻은 Root Mean Square(RMS) 에러는 0.95mm로 여타 다른 연구들에 비해 매핑의 정확도가 상대적으로 높은 편이었다. [12]에서는 해스킨스 연구소의 조음합성기인 TADA로부터 합성된 인공 조음과 음향 데이터를 사용하였다. 조음합성기의 데이터를 사용하여 훈련된 ANN 모델은 실제 발화 데이터인 Aurora-2 코퍼스를 사용한 테스트에서도 상당한 인식률을 보였다. [13]에서는 정밀한 물리학적 계산으로 합성해 낸 성도 및 기타 조음기관의 위치 값과 음성 시그널 간의 모델링을 이용한 speech inversion 모델의 훈련이 이루어졌다. 이를 위해 다양한 기계 학습 방식(trajectory mixture density networks(TMDNs), feedforward artificial neural networks(FF-ANN), support vector regression(SVR), autoregressive artificial neural network(AR-ANN), and distal supervised learning (DSL))이 적용되었다.

다양한 인공신경망의 적용은 기존 speech inversion에 존재하는 여러 제약을 개선할 수 있는 방안을 제공해 주었다. 그럼에도 불구하고, 과거 연구들 또한 일정한 한계가 존재했다. 이들 연구들은, 음성 발화와 동시에 기록된 실제 조음 데이터를 사용하되, 특정 음소들에만 국한된 speech inversion 모델을 제시하였다. 즉, 기존 연구들은 보편적인 조음데이터 산출에 내재적 한계점을 갖고 있다.

중요한 점은 이 문제가 인공신경망 자체의 한계에서 기인한다는 것이다. 인공 신경망의 연산 능력은 은닉층의 크기를 증가시키거나, 은닉층의 개수를 증가시키는 방식으로 개선될 수 있다. 하지만, 은닉층의 크기를 증가시킬 경우 과적합(over-fitting) 현상, 은닉층의 개수를 증가시킬 경우 정보의 손실로 인해 학습이 정상적으로 이루어지지 않는 현상이 발생할 여지가 있다. 따라서, 일정 수준 이상의 연산능력을 갖는데 제약이 발생하게 된다 [14].

이와 같은 인공신경망의 제약을 해결하고자, [15]에서는 Smolensky의 제한된 볼츠만 기계(restricted Boltzmann machine 이하 RBM)를 사전학습 과정(pre-training)으로 사용하는 deep belief network(DBN)를 제안하였다. 각 층 사이에 존재하는 가중치들을 무작위 값으로 두고 학습을 시작하는 기존 신경망 알고리즘과는 달리, DBN은 제한된 볼츠만 기계 알고리즘을 통해 가중치의 초기값을 얻는다. 이와 같은 방식의 사전학습 과정은 이후 학습이 정상적으로 이뤄질 수 있도록 신경망에 방향성을 부여한다. 결과적으로, DBN 알고리즘을 적용함으로써, 신경망은 은닉층이 다층으로 복잡하게 구성된 ANN에서 자주 발생하는 지역 최소점(local minima)에 빠지는 문제를 극복하고 개선된 학습을 진행해 나갈 수 있도록 한다.

Speech inversion 연구의 관점에서 볼 경우에도, 이와 같은 DBN 알고리즘의 등장은 중요한 의미를 지닌다. 위에서 언급한 바와 같이 speech inversion에 신경망을 적용한 선행연구들의 주된 제한점은 신경망의 연산 능력의 제약에서 기인하였다. 따라서 speech inversion에 DBN 알고리즘 적용가능성을 검증하는 것은 과거 많은 선행연구들의 제한점을 개선할 수 있는 한 가지 대안이 될 수 있을 것이다.

본 연구는 일반적인 ANN과 DBN 알고리즘을 적용한 심층신경망 모델(이하 DBN-DNN)을 구성하고, 각 모델에 speech inversion 정보를 학습시킴으로써 모델의 수행능력을 검증하고자 한다. 또한, 이를 통하여 해당 신경망 모델의 보편적인 조음 데이터 산출 가능여부를 검증해 봄으로써, 과거 선행연구들에서 포착된 난제를 극복해 보고자 한다.

2. 연구 방법 및 내용

2.1. 모델 생성을 위한 데이터

모델 생성에 사용된 데이터는 미국 Wisconsin대학에서 수집된 XRMB다 [16]. XRMB는 1989년에 처음 데이터베이스 구축에 대한 논의가 이루어졌고, 다년간의 데이터수집 후 1994년에 무료로 공개되었다. 데이터베이스는 57명 화자의 101개 발화 테스크 녹음 데이터(음소 + 조음)로 구성되어 있다. 발화를 수행하는 동안, x-ray microbeam이 피실험자들의 주요 조음기관들 여섯 군데, 즉 윗입술(upper lip 이하 UL), 아랫입술(lower lip 이하 LL), 혓끝(tongue tip 이하 T1), 혓날(tongue blade 이하 T2), 혓몸(tongue dorsum 이하 T3), 혓뿌리(tongue root 이하 T4)와 두 군

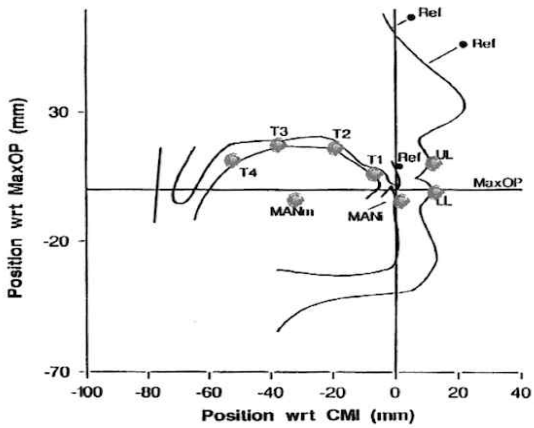

그림 1. XRMB의 6개의 주요 조음기관과 송곳니, 어금니의 위치
Figure 1. The locations of the 6 primary articulatory organs, molar, and incisor.

데의 기준점인 송곳니(mandibular incisior 이하 MANI), 어금니(mandibular molar 이하 MANM)에 부착된 금속 구슬들의 위치를 추적하고 기록하는 동시에 음성 또한 <그림 1>과 같이 녹음되었다. 음향에 해당하는 음성 데이터는 21,739 Hz 의 샘플링 주파수(sampling rate), 조음 데이터에 해당하는 여덟 기관들의 x축, y축 좌표 정보는 160 Hz의 샘플링 주파수로 기록되었다. 그림의 원점은 레퍼런스 위치가 되는 점으로 윗쪽 앞니의 중앙끝을 이용한다. 따라서 수평축에 대해서 목구멍 안쪽으로 음수의 좌표를 갖게 된다.

이 데이터베이스가 갖는 의의는 다음과 같다. 첫째, 음향과 조음 데이터를 동시에 수집하여 음향에서는 발견하기 힘든 조음 정보를 확인할 수 있도록 했다는 점이다. 음향데이터를 모으면서 이와 동시에 이루어지는 조음 데이터를 모으고 동기화하는 작업은 기술적으로 매우 어려운 부분인데, 최신 설비와 투자를 통해 이러한 데이터수집이 가능할 수 있었다. 둘째, 샘플 사이즈가 크고 다양한 언어적, 비언어적 태스크가 포함되어 있어 통계적 모델링과 기계학습을 위한 데이터로서 규모가 적절하다. 총 57명의 화자로부터 데이터를 수집하였고, 지문 읽기 등의 언어적 태스크뿐만 아니라 순수하게 실험 목적을 위한 비언어적 발화 태스크도 포함하고 있다. 모든 태스크의 총 녹음 시간은 약 20시간에 달한다. 셋째, 데이터베이스가 무료로 공개되었다는 점이다. 데이터 수집의 어려움과 비용 문제에도 불구하고 speech inversion을 포함한 여러 응용 분야의 연구와 발전을 위하여 처음부터 공개 배포를 목적으로 만들어졌다. 여타 조음 코퍼스는 특정 연구 목적으로 소규모로 수집되거나 비공개 데이터임에 반해, XRMB는 충분한 샘플 사이즈와 샘플의 다양성을 지녔을 뿐만 아니라 공개된 자료이므로 본 연구에 가장 부합한 자료로 판단되었다.

이번 연구에서는 학습 시간의 감소를 위해서 XRMB 데이터 중 무작위로 선정한 한 명의 화자의 데이터를 이용하였다. 선정된 데이터는 남성 화자의 데이터였으며 총 16,202개의 음향과 조음 데이터 쌍을 추출하였다. 음향 데이터는 발화 문장으로부터 13개의 필터뱅크(filterbank)를 계수(coefficients)로 하는 MFCC(Mel-frequency cepstral coefficients)를 구한 뒤, 그 값에서 각각 1차(13개), 2차(13개) 미분한 총 39개의 값을 수집하였다. 단, 음향데이터는 DBN-DNN에 적용하기 위하여, 해당 범주의 최소값을 빼고 0에서 1사이의 값으로 리스케일링 하였다. 다시 말해서, 입력층(input layer)에는 항상 0에서 1 사이의 값이 들어갈 수 있도록 조정되었다. 이에 대한 수학적 공식은 <수식 1>과 같다.

$$x'_i = \frac{x_i - \min_x}{\max_x - \min_x} \quad (1)$$

조음 데이터는 피험자의 주요 여덟 개의 조음기관들에 부착된 금속 구슬들의 16개의 위치 정보 (x,y축) 값으로 구성되었다. 학습용 데이터 쌍으로는 총 16,202개의 데이터 쌍으로부터 무작위 추출을 시행하여 13,000개를 선정하였으며, 중복되지 않은 나머지 3,202개의 학습 데이터 쌍을 테스트용의 데이터 쌍으로 사용하였다.

2.2. ANN 모델 생성

2.2.1. ANN 모델의 구조

ANN의 네트워크는 입력층(input layer), 출력층(output layer), 그리고 은닉층으로(hidden layer)으로 이루어져 있다. 각각의 층은 여러 개의 유닛으로 구성되어 있으며, 각 유닛은 개별적인 활성화(activation) 값을 갖는다. 또한 한 층의 모든 유닛들은 인접한 다른 층의 모든 유닛들과 연결되어 있고, 각각의 연결은 가중치(weight)를 갖고 있다. ANN의 학습은 입력값과 목표값(target)을

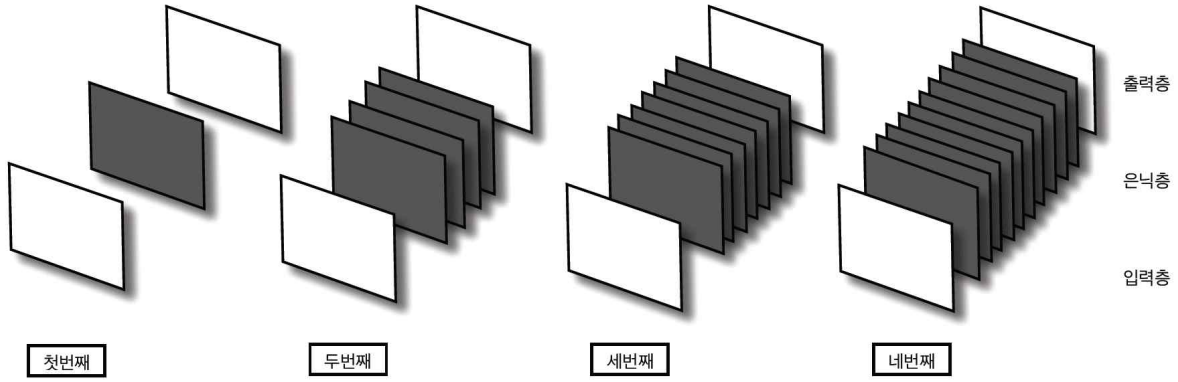

그림 2. speech inversion을 위한 네트워크 구성
Figure 2. The structure of the speech inversion network

쌍으로 준비하고, 입력값에 의해 도출된 출력값과 목표값의 차이인 오차를 계산, 이 오차를 점차적으로 줄여나갈 수 있도록 가중치를 갱신(update)하는 방식으로 진행된다. 이러한 학습은 오차 역전파(error back propagation) 알고리즘을 사용하여 진행되며, 목표값(target)과 출력값(output)의 오차(error)는 비용함수(cost function)를 이용하여 계산한다. <수식 2>는 본 연구에서 사용한 비용함수로써, 제곱 평균 제곱근(root mean square 이하 RMS)를 구하는 비용함수이다.

$$y_{rms} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{y_i}]^2} \quad (2)$$

y_i는 목표값을, $\hat{y_i}$ 실제 모델의 출력층에서 산출된 각 유닛의 활성화값을 의미한다. RMS의 물리적 크기는 사용된 데이터에 의해 결정되며, XRMB 데이터를 사용한 본 연구에서는 mm 단위의 물리적 크기를 갖는다.

본 연구에서는 ANN과 DNN의 성능을 비교하고자 하였으므로, 각각 은닉층의 개수를 1, 4, 7, 10개로, 각 은닉층의 유닛을 50, 100, 150개까지 달리한 총 12개의 ANN 모델을 설계했다. 입력층(input layer)과 출력층(output layer)은 XRMB 데이터에 맞춰, 각각 39개와 16개의 유닛으로 이루어졌다. <그림 2>는 은닉층 개수에 따라 모델이 어떻게 구성되었는지를 간략하게 나타낸다.

2.2.2. ANN 모델의 학습

ANN은 feedforward computation과 오차역전파(back propagation) 2단계로 진행되었다. FNN 과정은 입력층의 활성화 값을 다음 층으로 전달하여, 출력층의 활성화 값을 도출해내는 과정이다.

본 연구에서는 각 은닉층의 활성화 값 계산을 위하여 두 종류의 함수를 사용하였으며. 이에 따라 ANN은 활성화 값 계산 방식에 의해 ANN-Sigmoid와 ANN-ReLU로 나누어진다. 미세 조정 단계에서 시그모이드 함수(sigmoid function) DBN-DNN과의 면밀한 비교를 위해 ANN-Sigmoid 모델에 사용된 sigmoid 활성화 함수에도 0.9의 모멘텀 값이 적용되었으며, ANN-ReLU 모델은 ReLU 활성화 함수가 적용된 모델로, 기존 연구의 탁월한 학습 결과[16] 에 따라 본 연구에 사용되는 기계학습 네트워크에 적용하였다.

다만 두 모델의 출력층에서는 조음기관의 위치 값이 목표 값으로 설정되어 있으므로, 선형적인(linear) 활성화 함수를 이용하여 출력 값을 도출하였다.

FNN 과정이 끝나고 진행되는 오차역전파 과정은, 출력 값과 목표 값의 차이를 비용함수를 통해 계산하였고, 해당 학습의 오류를 산정하여 각 층간의 연결된 가중치와 바이어스 값을 수정해 나갔다. 본 연구에서는 총 100 세대(epoch)의 학습이 진행되었으며, 패턴들의 학습 순서에 의해 발생할 수 있는 편향을 제거하기 위해 각 세대별로 데이터 쌍들의 순서를 섞어주었다. 또한 학습률(learning rate)은 0.001 설정되었으며 이는 본 학습에 앞서 적절한 학습률을 구하고자 진행한 예비실험에서 가장 안정적인 학습이 진행되는 학습률 값으로 설정한 것이다. 예비실험에서 학습률은 0.1부터 시작하였으며, 추후 1/10만큼 차감해 나가면서 훈련결과의 과정을 비교하였다.

2.3. DBN-DNN 모델 생성

2.3.1. DBN-DNN 모델의 구조

DBN-DNN 의 경우, ANN과의 차이는 학습 과정에서 나타나므로, 구조상으론 ANN과 동일하다. 따라서 본 연구에서는, ANN의 구조와 마찬가지로 은닉층의 개수를 1, 4, 7, 10개, 은닉층의 유닛을 50, 100, 150개까지 달리하여 총 12개의 모델을 구축하였다. 이처럼 두 모델의 구조를 동일하게 설정함으로써, 두 모델 간 수행 차이에 나타날 수 있는 혼입변인을 배제시키고 오로지 두 모델간의 성능차이만 비교하고자 한다.

2.3.2. DNN 모델의 사전학습(pre-training)

ANN과 달리, DBN-DNN은 사전학습(pre-training)과 미세조정(fine-tuning)의 두 단계에 걸쳐 학습이 진행된다. 서론에서 언급

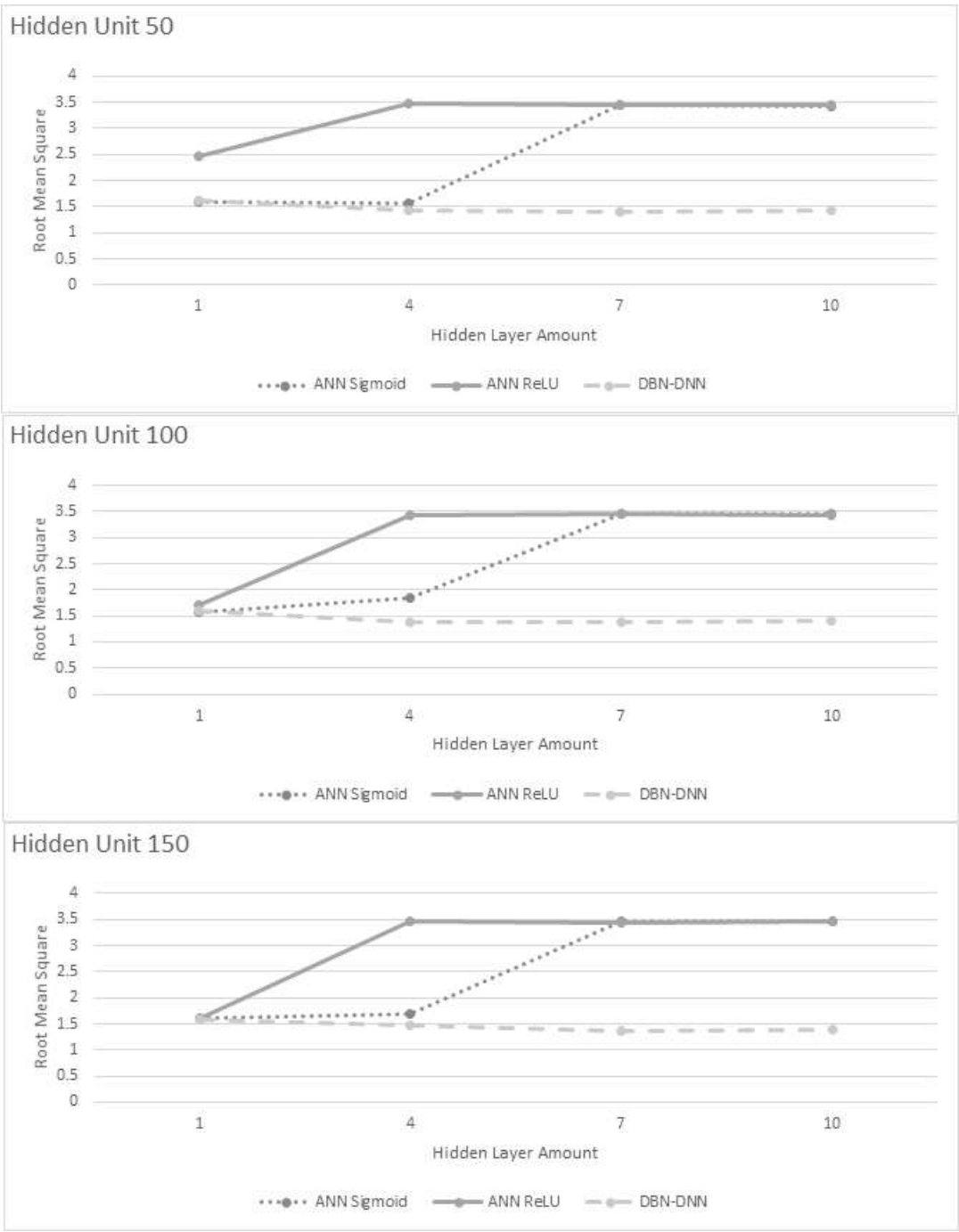

그림 3. 학습 종료 후 은닉층의 크기와 개수에 따른 RMS의 차이
Figure 3. The RMS results form three algorithms based on the sizes and the numbers of hidden layers.

한 바와 같이 DBN에서의 사전학습은 제한된 볼츠만 기계(RBM) 알고리즘을 greedy layer-wise training 방식으로 진행된다. 이 학습 방식은 하위층부터 최상위층까지 한 번에 학습이 진행되는 것이 아니라, 각 층에서 학습을 반복하여 진행한 뒤, 학습이 완료되면 다음 층으로 이동하여 전 층에서 시행한 횟수와 동일하게 반복 학습을 진행한다. 이 때 다음 층의 학습은 학습이 완료된 이전 층에서의 출력값을 이용하여 진행된다. 본 연구에서 각 DBN 모델은 층별로 10번의 학습 횟수와 0.1의 학습률을 사용하여 진행하였다.

또한 RBM의 사전학습에서 사용된 데이터는 추후 미세조정 학습에서 사용될 데이터 쌍 중 입력 값만을 가지고 진행되었으며, RBM 알고리즘의 비지도 학습(unsupervised learning) 특성상 목표 값은 이용하지 않았다.

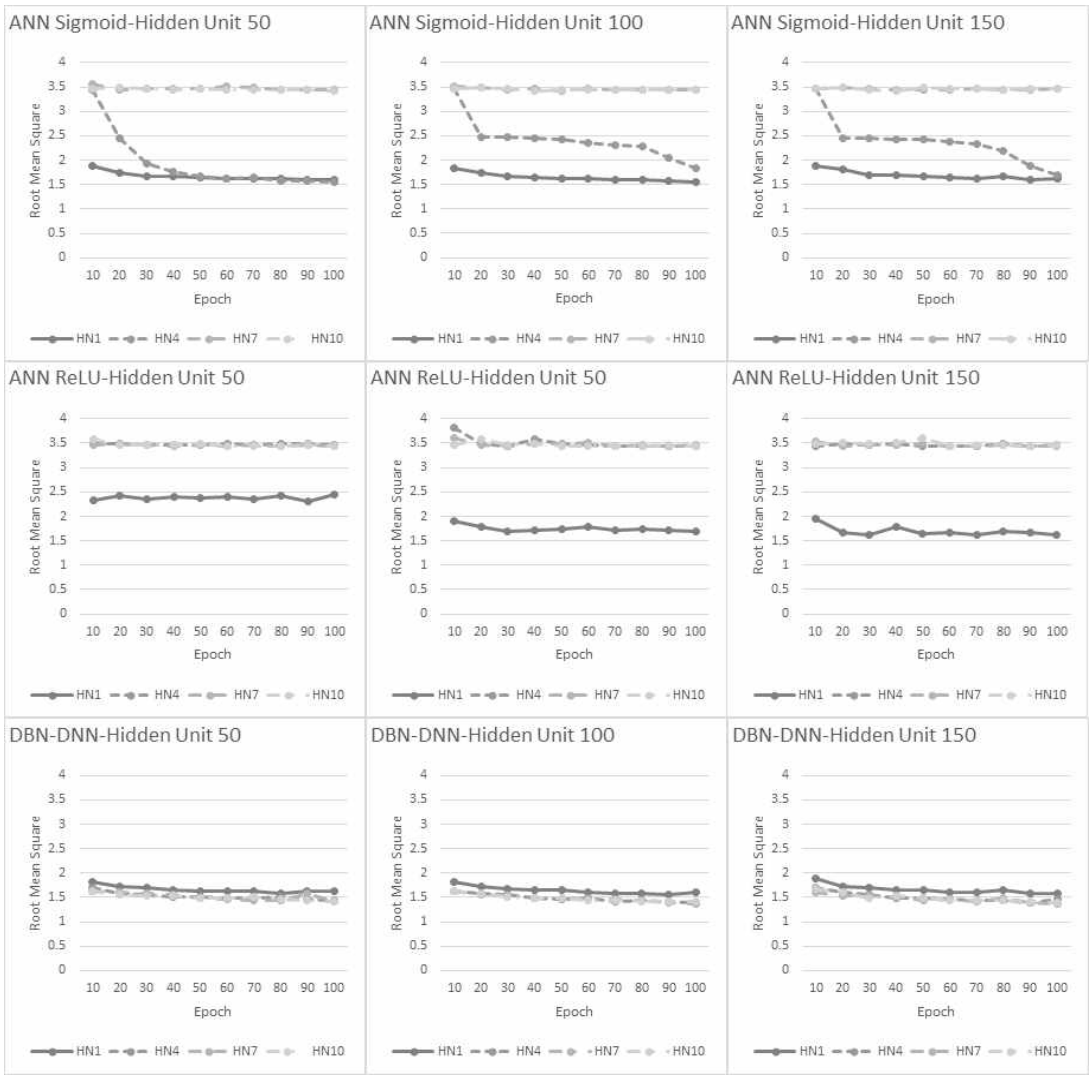

그림 4. 각 세대별 RMS의 변화. HN은 은닉층의 개수를 의미함.
Figure 4. The change of RMS in different numbers of layers. 'HN' means the numbers of hidden layers.

2.3.3. DBN-DNN 모델의 미세조정(fine-tuning)

DBN-DNN에서 미세조정 알고리즘은 복수로 존재하지만 본 연구에서는, 두 알고리즘 간의 비교를 명확하게 하기 위하여, ANN-Sigmoid의 학습과정에서 사용된 것과 동일한 오류역전파 알고리즘을 사용하였다. 따라서 ANN과 마찬가지로 총 100 세대(epoch)의 학습이 0.001의 학습률(learning rate)로 설정되어 진행되었으며, 각 세대별로 데이터 쌍들의 순서를 섞어주었다.

2.4. 훈련 모델 테스트

각 모델들은 학습하기 전 시점인 0세대를 포함하여 매 10세대의 학습이 진행될 때마다 테스트를 진행하였다. 이를 위해 미리 구성된 3,202개의 테스트용 데이터 셋이 사용되었으며, 비용함수를 이용하여 각 데이터 셋에 대한 RMS를 구하였다. 이를 통해 각 세대별로 모델들의 수행이 어떻게 변화하는지를 파악하며, 각 모델들의 최종적인 수행 능력의 차이를 확인하고자 하였다.

3. 결과 및 논의

3.1. 은닉층 개수 별 RMS 변화

<그림 3>은 학습이 종료된 후, 각 모델들의 RMS를 나타낸다. <수식 2>를 토대로 최소화된 에러 값을 구하는 것이 인공신경망의 목적인 점을 감안했을 때, 각 수치는 낮을수록 정확한 speech inversion이 가능했음을 의미한다.

<그림 3>에서 맨 위쪽은 각 은닉층의 크기인 은닉 유닛이 50인 모델들의 은닉층 개수에 따른 RMS의 변화량을 나타낸다. 특이한 점은 ReLU 함수를 사용한 ANN 모델이 Sigmoid 함수를 사용한 ANN모델보다 전반적으로 더 높은 RMS를 보여준다는 점인데, 이는 기존 연구에서 밝혀진 사실과 달리 speech inversion 문제를 훈련할 때는 ReLU 함수가 제대로 된 학습결과를 보여주

지 못한다는 것이다. ANN-Sigmoid 모델은 은닉층 개수가 4개인 지점까지 좋은 성능을 유지하였으나 그 이상의 증가에선 급격하게 하락하였다. 반면 DBN-DNN 모델은 모델의 은닉층 개수의 변화와 상관없이 RMS를 1.5 이하로 유지하였고, 오히려 은닉층의 개수가 증가함에 따라 성능이 향상되는 결과를 보여주었다.

<그림 3>의 중간에는 은닉 유닛이 100개인 모델들의 RMS 변화량을 나타낸다. 은닉 유닛이 50개인 위쪽 그림 조건과 비교하였을 때, 은닉층이 1개인 조건에서 ANN-ReLU 모델의 성능이 다른 모델들과 비슷하였으나, 은닉층의 개수가 증가하자 급격한 성능하락을 보였다. 반면 ANN-Sigmoid는 은닉층이 4개인 조건에서 약간의 성능하락이 있었으나, 다른 조건에서는 은닉 유닛이 50개인 경우와 큰 차이를 보여주지 않았다. DBN-DNN 모델은 은닉 유닛이 50개와 100개인 조건간의 차이가 없었다.

<그림 3>의 아래쪽은 은닉 유닛이 150개인 모델들의 RMS 변화량을 나타낸다. 이전의 두 조건과 비교하였을 때, 세 모델 모두 큰 변동이 없었으며, DBN-DNN 모델만이 약간의 성능향상을 보였다(은닉층이 10개이면서 각 은닉 유닛이 100개인 조건간의 평균 RMS 차이는 약 0.013이다).

또한 절대적인 수행능력 측면을 비교해 보았을 때, ANN-Sigmoid(은닉 유닛: 50개, 은닉층: 4개인 모델의 RMS: 1.551)와 ANN-ReLU(은닉 유닛: 150개, 은닉층: 1개인 모델의 RMS: 1.616)의 최고 훈련 성능이, DBN-DNN의 최고 훈련 성능(은닉 유닛: 150개, 은닉층: 7개인 모델의 RMS: 1.370)보다 낮음을 보여주었다. 이와 같은 결과는 ANN모델들에 비해 DBN-DNN 모델이 안정성을 가지고 조음 데이터를 학습한다고 볼 수 있다.

3.2. 세대 별 RMS 변화

<그림 4>의 맨 위쪽 세 그래프는 ANN-Sigmoid 모델들의각 세대 별 RMS를 나타낸다. 은닉층이 1개인 조건에서는 맨 밑의 DBN-DNN 그래프와 유사하게 RMS가 1.5 부근의 지점에서 어트랙터 상태(안정화 단계)에 들어감을 보여주는 반면, 은닉층이 4개로 늘어난 조건에서는 더 많은 세대의 학습이 요구되었으며, 7개의 조건에서는 정상적으로 학습이 이뤄지지 않고 있음을 보여준다.

<그림 4>의 중간의 세 그래프는 ANN-ReLU 모델들의 각 세대별 RMS를 나타낸다. 3.1에서 언급한 바와 같이 ANN-ReLU 모델들은 오직 은닉층이 1개인 조건에서만 어트랙터 상태에 도달하였으며, 은닉 유닛이 50개인 조건의 경우, 은닉 유닛이 100개와 150개인 조건보다 RMS가 높게 나타났다.

반면 각 세대 별 DBN-DNN 모델은 ANN 모델과는 다르게 은닉층의 개수가 증가하여 네트워크가 복잡해짐에도 불구하고 RMS가 점차 낮아지거나 유지되면서 올바른 학습 진행 양상을 보여준다.

4. 결론

본 연구는 일반적인 인공신경망(ANN)과 deep belief network(DBN) 알고리즘을 적용한 심층신경망 모델(DNN)을 구성하고, 각 모델에 speech inversion 정보를 학습시킴으로써 모델의 수행능력을 검증해 보고자 하였다. 또한, 이를 통하여 각 신경망 모델의 보편적인 조음 데이터 산출 가능여부를 검증해 봄으로써, 과거 선행연구들에서 나타난 한계점이 현실적으로 극복가능한지 확인해보고자 하였다.

본 연구의 결과는 종래의 ANN과는 다르게 DBN-DNN이 다층구조의 은닉층을 보유하였음에도 정상적으로 데이터를 훈련할 수 있음을 보였고, 나아가 전반적으로 더 나은 수행을 보일 수 있었다는 점에서 큰 의미를 지닌다. 조음 데이터는 데이터 확보에 많은 시간과 노력이 들기 때문에 대규모의 데이터를 구축하는데 큰 제한이 따른다. 이런 상황 속에서, 소규모 데이터로도 효율적인 학습을 유도하는 DBN-DNN 알고리즘을 speech inversion에 적용할 수 있다는 가능성은, 차후 speech inversion 관련 연구 및 다양한 실용 분야에서 적용될 것으로 기대된다.

참고문헌

[1] Ghosh, P. K. & Narayanan, S. (2011). Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion. *The Journal of the Acoustical Society of America*, 130(4), EL251-EL257.

[2] Sondhi, M. M. & Resnick, J. R. (1983). The inverse problem for the vocal tract: Numerical methods, acoustical experiments, and speech synthesis. *The Journal of the Acoustical Society of America*, 73(3), 985-1002.

[3] Wilson, I., Gick, B., O'Brien, M. G., Shea, C., & Archibald, J. (2006). Ultrasound technology and second language acquisition research. *Proceedings of the 8th Generative Approaches to Second Language Acquisition Conference* (GASLA 2006) (pp. 148-152).

[4] Wrench, A. A., Gibbon, F., McNeill, A. M., & Wood, S. (2002). An EPG therapy protocol for remediation and assessment of articulation disorders. *ICSLP*.

[5] Dusan, S. (2001). Methods for integrating phonetic and phonological knowledge in speech inversion. *Proceedings of the International Conference on Speech, Signal and Image Processing*. Malta.

[6] Engwall, O. (2006). Evaluation of speech inversion using an articulatory classifier. *Proceedings of the 7th International Seminar on Speech Production* (pp. 469-476).

[7] Papcun, G., Hochberg, J., Thomas, T. R., Laroche, F., Zacks, J., & Levy, S. (1992). Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data. *The Journal of the Acoustical Society of America*, 92(2), 688-700.

[8] Zacks, J. & Thomas, T. R. (1994). A new neural network for articulatory speech recognition and its application to vowel identification. *Computer Speech & Language*, 8(3), 189-209.

[9] Richmond, K. (2001). Mixture density networks, human articulatory data and acoustic-to-articulatory inversion of continuous speech. *Proceedings of Workshop on Innovation in Speech Processing* (WISP 2001) (pp. 259-276).

[10] Qin, C. & Carreira-Perpinán, M. A. (2010). Articulatory inversion of american english /r/ by conditional density modes. *Proceedings of 11th Annual Conference of the International Speech Communication Association* (Interspeech 2010) (pp. 1998-2001)

[11] Richmond, K., Hoole, P., & King, S. (2011). Announcing the Electromagnetic Articulography (Day 1) Subset of the mngu0 Articulatory Corpus. *Proceedings of 12th Annual Conference of the International Speech Communication Association* (Interspeech 2011) (pp. 1505-1508).

[12] Mitra, V., Nam, H., Espy-Wilson, C., Saltzman, E., & Goldstein, L. (2011). Articulatory information for noise robust speech recognition. Audio, Speech, and Language Processing, *IEEE Transaction on Audio, Speech, and Language Processing*, 19(7), 1913-1924.

[13] Najnin, S. & Banerjee, B. (2015). Improved speech inversion using general regression neural network. *The Journal of the Acoustical Society of America*,138(3), EL229-EL235.

[14] Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of clinical epidemiology*, 49(11), 1225-1231.

[15] Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, 18(7), 1527-1554.

[16] Simpson, A. J. (2015). Taming the ReLU with Parallel Dither in a Deep Neural Network (arXiv preprint). Retrieved from http://arxiv.org/abs/1509.05173 on September 17, 2015

• **유희조 (You, Heejo)**
고려대학교 심리학과
서울시 성북구 안암로 145
Email: codejin@korea.ac.kr
관심분야: 언어심리, 언어모델링

• **양형원 (Yang, Hyungwon)**
고려대학교 영어영문학과
서울시 성북구 안암로 145
Email: hyung8758@korea.ac.kr
관심분야: 음성학, 언어공학

• **강재구 (Kang, Jaekoo)**
고려대학교 영어영문학과
서울시 성북구 안암로 145
Email: zzandore@korea.ac.kr
관심분야: 음성학, 언어공학

• **조영선 (Cho, Youngsun)**
고려대학교 영어영문학과
서울시 성북구 안암로 145
Email: youngsunhere@korea.ac.kr
관심분야: 음성학, 언어공학

• **황성하 (Hwang, Sung Hah)**
고려대학교 영어영문학과
서울시 성북구 안암로 145
Email: hshsun@korea.ac.kr
관심분야: 음성학, 언어공학

• **홍연정 (Hong, Yeonjung)**
고려대학교 영어영문학과
서울시 성북구 안암로 145
Email: yvonne_yj_hong@korea.ac.kr
관심분야: 음성학, 언어공학

• **조예진 (Cho, Yejin)**
고려대학교 영어영문학과
서울시 성북구 안암로 145
Email: scarletcho@korea.ac.kr
관심분야: 음성학, 언어공학

• **김서현 (Kim, Seohyun)**
고려대학교 영어영문학과
서울시 성북구 안암로 145
Email: sh77@korea.ac.kr
관심분야: 음성학, 언어공학

• **남호성 (Nam, Hosung)** 교신저자
고려대학교 영어영문학과
서울시 성북구 안암로 145
Tel: 02-3290-1991
Email: hnam@korea.ac.kr
관심분야: 음성학, 언어공학