한국음향학회지 제38권 제6호 pp. 703~709 (2019)
*The Journal of the Acoustical Society of Korea* Vol.38, No.6 (2019)
https://doi.org/10.7776/ASK.2019.38.6.703

pISSN : 1225-4428
eISSN : 2287-3775

# 심층신경망 구조에 따른 구개인두부전증 환자 음성 인식 향상 연구

# A study on recognition improvement of velopharyngeal insufficiency patient's speech using various types of deep neural network

김민석,[1] 정재희,[1] 정보경,[1] 윤기무,[1] 배아라,[1] 김우일[1†]

(Min-seok Kim,[1] Jae-hee Jung,[1] Bo-kyung Jung,[1] Ki-mu Yoon,[1] Ara Bae,[1] and Wooil Kim[1†])

[1]인천대학교 컴퓨터공학부

(Received October 8, 2019; accepted October 29, 2019)

**초 록**: 본 논문에서는 구개인두부전증(VeloPharyngeal Insufficiency, VPI) 환자의 음성을 효과적으로 인식하기 위해 컨볼루션 신경망(Convolutional Neural Network, CNN), 장단기 모델(Long Short Term Memory, LSTM) 구조 신경망을 은닉 마르코프 모델(Hidden Markov Model, HMM)과 결합한 하이브리드 구조의 음성 인식 시스템을 구축하고 모델 적응 기법을 적용하여, 기존 Gaussian Mixture Model(GMM-HMM), 완전 연결형 Deep Neural Network(DNN-HMM) 기반의 음성 인식 시스템과 성능을 비교한다. 정상인 화자가 PBW452단어를 발화한 데이터를 이용하여 초기 모델을 학습하고 정상인 화자의 VPI 모의 음성을 이용하여 화자 적응의 사전 모델을 생성한 후에 VPI 환자들의 음성으로 추가 적응 학습을 진행한다. VPI환자의 화자 적응 시에 CNN-HMM 기반 모델에서는 일부층만 적응 학습하고, LSTM-HMM 기반 모델의 경우에는 드롭 아웃 규제기법을 적용하여 성능을 관찰한 결과 기존 완전 연결형 DNN-HMM 인식기보다 3.68 % 향상된 음성 인식 성능을 나타낸다. 이러한 결과는 본 논문에서 제안하는 LSTM-HMM 기반의 하이브리드 음성 인식 기법이 많은 데이터를 확보하기 어려운 VPI 환자 음성에 대해 보다 향상된 인식률의 음성 인식 시스템을 구축하는데 효과적임을 입증한다.

**핵심용어**: 구개인두부전증(VeloPharyngeal Insufficiency, VPI), 음성인식, 컨볼루션 신경망, 장단기 메모리, 심층 신경망

**ABSTRACT:** This paper proposes speech recognition systems employing Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) structures combined with Hidden Markov Moldel (HMM) to effectively recognize the speech of VeloPharyngeal Insufficiency (VPI) patients, and compares the recognition performance of the systems to the Gaussian Mixture Model (GMM-HMM) and fully-connected Deep Neural Network (DNN-HMM) based speech recognition systems. In this paper, the initial model is trained using normal speakers' speech and simulated VPI speech is used for generating a prior model for speaker adaptation. For VPI speaker adaptation, selected layers are trained in the CNN-HMM based model, and dropout regulatory technique is applied in the LSTM-HMM based model, showing 3.68 % improvement in recognition accuracy. The experimental results demonstrate that the proposed LSTM-HMM-based speech recognition system is effective for VPI speech with small-sized speech data, compared to conventional GMM-HMM and fully-connected DNN-HMM system.

**Keywords:** VeloPharyngeal Insufficiency (VPI), Speech recognition, Convolutional neural network, Long short term memory, Deep neural network

**PACS numbers:** 43.72.Bs, 43.72.Ne

†**Corresponding author:** Wooil Kim (wikim@inu.ac.kr)
Department of Computer Science and Engineering, Incheon National University, 119 Academy-ro, Yeonsu-gu, Incheon 22012, Republic of Korea
(Tel: 82-32-835-8459, Fax: 82-32-835-0780)

## I. 서 론

구순구개열은 얼굴의 선천성 기형 중 빈도가 높은 장애의 하나로, 선천적으로 입술(구순) 또는 입천장(구개)이 갈라져서 구강과 비강이 연결된 상태를 말하며, 우리나라의 경우 약 700명의 신생아 중 1명 꼴로 발생하는 유병률이 높은 질환이다.[1,2] 정상인이 음성을 발음할 때는 연구개가 비강과 구강을 차단함으로써 비음을 제한하지만, 구순구개열 환자의 경우에는 경구개 또는 연구개가 벌어져 있거나 선천적으로 연구개가 짧기 때문에, 발성 시에 성대로부터 나온 기류가 비강과 구강을 동시에 공명하게 되어 발성 및 조음 장애를 일으키게 된다. 이러한 증상을 구개인두부전증(VeloPharyngeal Insufficiency, VPI)이라고 한다.

본 논문은 VPI 환자의 음성을 정상인의 음성에 가깝게 복원하고 향상시켜, 정확하게 자동으로 인식하기 위한 기법 연구에 관한 것이다. VPI 환자 음성을 정확하게 자동으로 인식하기 위한 기법에 관한 연구결과를 소개한다. 본 연구의 사전 연구로서, VPI 환자의 음성에 관한 효과적인 연구를 위해 PBW452 데이터베이스의 단어 목록으로 VPI환자에게 수집한 공동 음성 데이터베이스를 구축하고, VPI 환자의 실제 발음과 정상인으로부터 실험적으로 VPI 환자 음성과 유사하게 발생시킨 모의 음성의 분석을 실시하였다.[3] 정상인으로부터 수집한 모의 VPI 환자 음성을 화자 적응의 초기 모델로 사용함으로써 다량의 음성 데이터 수집이 어려운 제한 조건을 갖는 VPI 환자 음성 인식에서 성능 향상을 관찰하였다.[4] 완전 연결된 심층 신경망과 은닉 마르코프 모델의 하이브리드 구조의 음향 모델을 갖는 음성 인식 시스템을 구축하여 VPI 음성에 대한 음성 인식 성능을 평가하였다.[5]

본 논문에서는 그 후속 연구로서 컨볼루션 신경망(Convolutional Neural Network, CNN) 및 장단기 메모리(Long Short Term Memory, LSTM) 구조의 심층 신경망과 은닉 마르코프 모델의 하이브리드 구조의 음향 모델을 갖는 음성 인식 시스템을 구축하여 VPI 음성에 대한 인식 성능을 평가한다. 사전 연구와 동일한 과정으로 모의 VPI 음성 데이터를 이용하여 초기 모델을 생성하고 이를 기반으로 하여 CNN 및 LSTM 기반 하이브리드 모델 구조에 VPI 환자 음성에 대한 화자 적응을 적용했을 때 성능 향상을 관찰한다. CNN 구조에서는 컨볼루션 층, 풀링 층, 은닉 층의 개수 및 필터 사이즈 및 구조를 변경하여 그 성능을 관찰한다. 화자 적응 성능을 높이기 위해 신경망 모델 전체 또는 일부 층만을 선택적으로 학습하여 인식 성능을 관찰한다. LSTM 구조에서는 화자 적응 성능을 높이기 위해 드롭 아웃 규제 기법을 적용하여 그 성능을 관찰하고 비교한다.

## II. 음성 데이터베이스 수집

VPI 환자 음성 샘플 수집을 위한 발음 목록으로 한국어 음성 인식 분야에서 많은 연구자들이 사용하고 있는 PBW452 데이터베이스[3]의 단어 목록을 사용하였다. PBW452 단어 목록의 452개 단어 중에서 언어치료사가 선정한 VPI 환자의 발음 오류 빈도수가 높은 50개 단어를 발음 목록으로 사용했다. 수집 대상으로는 음성 녹음에 협조가 잘 이루어지고, 발음 목록에 따른 발성 과정이 적절히 수행될 수 있도록 만 10세 이상의 VPI 환자를 선정하였다. VPI 모의 음성 수집을 위해서는 정상 발음을 가진 성인을 대상으로 하였다. 모집과정에서 VPI 환자와 정상인 모의 발음 대상의 녹음 의지를 확인한 후 피험자 동의서를 받았다(IRB Number; 1103-040-354, [2]).

정상인이 실험적으로 VPI 환자의 발음과 유사하게 발음하기 위하여 1 mm 내경을 갖는 고무관(Nelaton Catheter)을 사용하였다.[3] 고무관을 양측 비강을 통해 삽입하고 긴장도가 없는 상태의 위치를 지혈겸자로 표시하고, 통증을 유발하지 않으면서 최대의 긴장도가 생성되는 위치를 표시하였다. 카테터가 최대 긴장도를 생성하는 위치에 있을 때를 VPI 모의 환자의 중증(severe) 상태로 정의하고, 최대 긴장도 위치와 긴장도가 없는 위치의 중간에 있을 때 화자의 발성이 녹음된 것을 경도(mild) 상태로 정의하였다. 본 논문에서는 중증 상태의 VPI 모의 환자 발음을 사용하였다.

## III. GMM-HMM 기반 음성 인식 및 모델 적응 기법

본 논문에서는 VPI 음성에 대해 Gaussian Mixture Model-Hidden Markov Model(GMM-HMM) 기반의 음성 인식 성능 평가를 위해 HTK[6] 소프트웨어와 PBW 452 음성 데이터베이스를 이용하여 기본 음성 인식 시스템을 구축하였다. 연구용으로 배포된 버전에 포함되어 있는 남자 8명이 2회씩 발음한 452단어 총 7,232개의 음성 샘플을 훈련 데이터로 이용하여 452개의 독립 단어를 인식하는 기본 음성 인식기를 구축하였다. 묵음 구간 모델을 포함하여 총 47개의 문맥 독립형 음소 모델을 사용하였다. 각 음소 모델은 하나의 HMM에 대응되며, 각 HMM은 3개의 상태(state)로 구성되고 각 상태는 출력 확률 함수로서 8개의 요소로 이루어진 가우시안 혼합 모델을 갖는다. ETSI 표준으로 정의한 Mel Frequency Cepstral Coefficients(MFCC) 특징 추출 기법[7]을 사용하여 c0를 포함한 13차 MFCC 특징(c0 ~ c12)에 1차, 2차 미분 계수를 결합하여 총 39차원의 특징 벡터를 추출하였다. 베이스라인 음성 인식 시스템의 성능 평가를 위해 훈련 데이터베이스 세트와 중복되지 않은 PBW452 평가용 버전의 남성 화자 5명의 깨끗한 발음 총 2,260 샘플을 사용하여 평가한 결과, 깨끗한 환경의 정상인 음성 데이터에 대해 98.89 %의 단어 인식률을 나타냈다.

음성 인식 성능 향상을 위한 음향 모델 적응 기법에서는 실제 인식 시스템이 적용되는 테스트 환경과 인식 시스템의 음향 모델 훈련이 이루어진 환경의 음향적 조건이 동일할 때 가장 높은 성능을 가지는 것을 가정한다. 본 연구에서는 GMM-HMM 기반 음성 인식을 위한 대표적인 모델 적응 기술인 Maximum A Posteriori(MAP) 기반 적응 기법[8]과 Maximum Likelihood Linear Regression(MLLR) 기반 적응 기법[9]을 사용하였다.

## IV. 다양한 심층신경망 기반 음성 인식 및 모델 적응 기법

본 논문에서는 GMM-HMM 기반의 VPI 음성 인식 성능과 비교하기 위해 다양한 구조의 DNN과 HMM의 하이브리드 구조를 갖는 음성 인식 시스템을 구축하였다.[10,11]

### 4.1 DNN-HMM 하이브리드 시스템

DNN-HMM 하이브리드 기반 음성 인식 시스템에서는 GMM-HMM 기반 음성 인식기와 동일한 구조의 HMM을 사용하며 HMM의 각 상태에 대한 확률 값을 GMM을 이용하여 계산하는 대신 DNN을 통해 출력되는 값을 확률 값으로 대체하여 사용한다. 확률값 대체를 위해 출력 단에서의 활성화 함수는 Softmax 함수를 사용한다. 입력 음성의 이웃한 5개 프레임에 대해 각 39차원 MFCC 특징 벡터를 결합한 총 195차원의 특징 벡터를 입력 데이터로 사용하였다. 이에 따라 입력 층의 뉴런의 개수는 음성 특징 벡터의 사이즈와 동일한 195개를 갖고, 출력 층의 뉴런의 개수는 모든 47개의 HMM 모델의 총 상태 개수와 동일한 141개를 사용하였다.

신경망 학습을 위해서는 우선 GMM-HMM 음성 인식기를 이용하여 훈련에 사용된 PBW 데이터를 forced-alignment를 통해 각 훈련 데이터에 대한 상태 열을 생성한다. 생성된 상태 열 정보가 DNN 학습 시에 레이블 데이터가 된다. 각 훈련 데이터에 대해 생성된 레이블 데이터를 목표로 하여 심층 신경망을 학습한다. 인식 시에는 GMM-HMM 인식기와 동일한 상태 천이 확률 모델을 사용하고, 심층 신경망 기반의 모델을 통하여 각 상태에 대한 확률 값을 계산하며 비터비 디코딩 과정은 GMM- HMM 기반 시스템과 동일하다.

### 4.2 CNN-HMM 하이브리드 시스템

L개의 컨볼루션(convolution) 층과 M개의 풀링(pooling) 층을 갖는 CNN 구조에서 $l$번째 컨볼루션 층의 $(i,j)$ 위치에서의 출력 값 $(f*g)^l(i,j)$은 다음과 같이 표현할 수 있다.[12,13]

$$(f*g)^l(i,j)=\sum_{x=0}^{h-1}\sum_{y=0}^{w-1}f^l(x,y)g^{l-1}(i+x,j+y). \quad (1)$$

위 식에서 함수 $f^l()$은 $l$번째 층에서 커널 함수의 행렬을 나타내며 $g^{l-1}()$는 $l-1$번째 층의 출력 데이터

행렬을 나타낸다. $h$와 $w$는 커널 함수의 사이즈를 나타낸다.

본 논문에서 제안한 CNN-HMM 기반의 하이브리드 시스템에서는 입력 음성의 이웃한 5개 프레임의 각 23차원 Log-spectrum 특징 벡터를 결합한 총 115차원의 특징 벡터를 입력 데이터로 사용하였다. 이에 따라 입력 층의 뉴런의 개수는 음성 특징 벡터의 사이즈와 동일한 115개를 갖고, 출력 층의 뉴런의 개수는 47개의 HMM 모델의 총 상태 개수와 동일한 141개를 사용하였다. VPI 환자 모델 적응을 위해 CNN 모델의 모든 층을 학습하는 것이 아닌 일부 층만을 학습하여 모델 적응을 실시하였다.

### 4.3 LSTM-HMM 하이브리드 시스템

L개의 LSTM(Long Short-Term Memory) 셀(cell)을 갖는 구조에서 시간 $t$ 에서의 셀의 상태 벡터 $c_t$, 망각 게이트 $f_t$, 입력 게이트 $i_t$, 출력 게이트 $o_t$, 출력 값 $h_t$는 다음과 같이 표현된다.[14,15]

$$f_t = \sigma(W_{xh\_f}x_t + W_{hh\_f}h_{t-1} + b_{h\_f}). \quad (2)$$

$$i_t = \sigma(W_{xh\_i}x_t + W_{hh\_i}h_{t-1} + b_{h\_i}). \quad (3)$$

$$o_t = \sigma(W_{xh\_o}x_t + W_{hh\_o}h_{t-1} + b_{h\_o}). \quad (4)$$

$$g_t = \tanh(W_{xh\_g}x_t + W_{hh\_g}h_{t-1} + b_{h\_g}). \quad (5)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t. \quad (6)$$

$$h_t = o_t \odot \tanh(c_t). \quad (7)$$

위 식에서 $\odot$는 벡터의 요소별 곱셈 연산을 나타내고, $W$와 $b$는 학습 과정에서 추정되는 weight와 바이어스 파라미터를 나타낸다. LSTM-HMM 하이브리드 시스템에서는 CNN-HMM 시스템과 동일한 115차원의 Log-spectrum 특징 벡터를 입력 데이터로 사용하였다. 모델 적응 과정에서는 모든 층을 학습하고 드롭 아웃 규제를 적용하였다.

## V. 실험 및 결과

사전 연구에서 제안한 GMM-HMM, DNN-HMM모델을 5.1과 5.2에 서술하였으며, 본 논문에서 제안한 CNN-HMM, LSTM-HMM모델과 비교 실험을 진행하였다.

### 5.1 GMM-HMM 기반 음성 인식 성능 평가

Table 1은 기존의 전통적 음성 인식 시스템 구조인 GMM-HMM 기반 음성 인식기를 이용하여 VPI 환자 음성에 대해 음성 인식 성능을 평가한 결과이다. 아무런 처리를 하지 않았을 때(no processing)는 환자 1과 2(P1, P2) 모두 2.68 %와 39.33 %로 낮은 인식률을 보였다. 소량의 VPI 음성 데이터 샘플을 화자 적응 실험에 효과적으로 이용하기 위해 각 화자의 샘플을 모델 적응용과 테스트용 세트가 겹치지 않도록 조합된 3종류의 세트로 구성하였다. 즉, 3회 반복 녹음으로 구성된 각 화자의 데이터에서 1회째 녹음 음성을 테스트할 때에는 2회째, 3회째 녹음 음성을 화자 적응에 사용하고, 2회째, 3회째 발음을 테스트 할 때에는 1회째, 3회째 발음과 1회째, 2회째 녹음 음성을 각각 화자 적응에 사용하였다. MLLR 화자 적응 실험 결과 두 환자에 대해 평균 84.06 %의 높은 인식률을 보였다. 이러한 결과는 VPI 환자 음성을 정확하게 인식하기 위해서는 대상 환자 음성을 이용한 화자 적응 과정이 필수적으로 요구됨을 의미하며, 실제 VPI 환자 음성과 같이 대량의 데이터 확보에 한계가 있을 경우에는 MLLR 기반의 음향 모델 적응 기법이 화자 적응에 효과적임을 입증한다.

Table 1에서 MAP-MLLR 실험은 정상인의 모의 음성을 이용하여 MAP 모델 적응을 통해 모의 음성 모델을 생성하고 이를 기반으로 MLLR 화자 적응을 적용하여 환자 1과 2(P1, P2) 음성에 대한 인식 성능 평

Table 1. GMM-HMM bases speech recognition results of speaker adaptation (word accuracy, %).

| | P1 | P2 | Avg. |
|---|---|---|---|
| No Processing | 2.68 | 39.33 | 21.01 |
| MLLR | 78.52 | 90.67 | 84.06 |
| **MAP-MLLR** | **85.23** | **92.67** | **88.95** |

가 결과를 나타낸다. 즉, PBW452 데이터를 이용하여 훈련된 기본 인식기를 초기 모델로 사용하고, 모의 음성을 이용하여 MAP 모델 적응을 실시함으로써 화자 적응을 위한 사전 모델을 생성하였다. VPI 모의 음성 모델을 생성하기 위한 모델 적응 과정에는 정상인 5명이 발음한 모의 발음 총 698개의 음성 데이터를 사용하였다. 모의 음성 모델을 화자 적응의 사전 모델로 이용함으로써 두 환자에 대해 인식률이 평균 88.95 %로 4.89 % 향상되었다.

## 5.2 DNN-HMM 기반 음성 인식 성능 평가

Table 2는 완전 연결 구조를 갖는 신경망을 사용한 하이브리드 구조(DNN-HMM) 기반의 음성 인식기를 이용하여 VPI 환자 음성에 대해 음성인식 평가를 시행한 결과이다. 일련의 다양한 실험을 거쳐 2개의 은닉 층이 각각 500개와 200개의 뉴런을 갖는 신경망 구조가 가장 높은 인식률을 보이는 것을 확인했다. 활성함수로는 각 층에서 ReLU 함수를 사용하고 출력층에서는 확률 형태의 값을 얻기 위하여 Softmax 함수를 사용하였다. GMM-HMM 학습과 동일한 PBW452 데이터를 이용하여 초기 모델을 학습하고 이를 기반으로 정상인의 모의 음성 데이터를 이용하여 학습했다. 이는 GMM-HMM 음성 인식기의 MAP 모델을 생성하는 과정과 유사한 과정이다. 두 단계 모두 학습률은 0.001을 사용하고 배치 사이즈는 학습 데이터 양의 10 % 크기로 하였다. Adam 최적화 알고리즘을 사용하여 교차 엔트로피 기준으로 학습하였다. 다양한 실험을 통해 최고의 성능을 보이는 파라미터 조합을 선택하고, 초기 학습과 화자 적응 단계에서는 각각 다른 학습률을 사용했다.

모의 음성 데이터를 이용하여 학습 후 얻은 DNN 모델을 기반으로 하여 Table 1의 MLLR 화자 적응 과정과 동일한 데이터 세트를 사용하여 화자 적응을 하였다. 화자 적응에서는 소량의 학습 데이터를 고려하여 각 가중치 행렬을 부분적으로 학습에 참여하여 인식 성능을 관찰했다. 가중 행렬 W1만 또는 W1와 W2만 화자 적응 학습에 참여했을 때 가장 높은 성능을 나타내었다. 화자 적응을 위한 학습에서는 0.0002의 학습률을 사용하고 배치 사이즈는 학습 데이터양의 25 % 크기로 했다.

## 5.3 CNN-HMM 기반 음성 인식 성능 평가

CNN-HMM 기반 하이브리드 구조의 음성 인식 시스템을 이용한 VPI 환자 음성 인식 성능을 평가하기 위해 다양한 구조를 갖는 CNN을 적용하여 실험을 진행하였다. 실험에서는 컨볼루션 층, 풀링 층, 은닉 층의 개수, 필터 사이즈 및 개수를 변경하여 성능을 평가하였다. 평가 결과, 본 실험에서는 4개의 컨볼루션 층, 2개의 풀링 층, 2개의 은닉 층으로 구성하고, 컨볼루션의 각 층은 3 × 3 커널 크기를 갖는 필터를 32개, 64개, 126개, 256개를 적용하였을 때 Table 3의 첫 번째 행과 같이 90.62 %로 가장 좋은 성능을 나타냈다. 화자 인식 성능 향상을 위해 일부 층만을 학습함으로써 성능 향상을 확인하였고, 본 논문에서는 컨볼루션 층 중 세 번째 층 C3, 네 번째 층 C4과 첫 번째 은닉 층 L1 만을 학습하여 가장 높은 성능을 얻을 수 있었다. CNN 구조의 층과 필터의 개수를 변경시키는 것이 다소 성능의 향상을 가져오지만, 인식 성능에 한계가 있고 학습시간이 증가하여 무조건 증가시키는 것이 효율적이지 못함을 알 수 있었다. 결과적으로 CNN-HMM 구조는 기존의 완전 연결형 DNN-HMM 구조와 비교하여 2.01 %의 인식률 향상을 보였다.

## 5.4 LSTM-HMM 기반 음성 인식 성능 평가

LSTM-HMM 기반 하이브리드 시스템의 성능을 평가하기 위해 뉴런 개수, 은닉층 수, 규제 기법 등을 조정하며 실험을 진행하였다. Table 4의 결과에서는

Table 2. DNN-HMM bases speech recognition results of speaker adaptation (word accuracy, %).

| Trained Weights | P1 | P2 | Avg. |
|---|---|---|---|
| All (W1, W2, W3) | 82.55 | 93.33 | 87.94 |
| **W1, W2** | **85.23** | **94.67** | **89.95** |

Table 3. CNN-HMM bases speech recognition results of speaker adaptation (word accuracy, %).

| Trained Layers | P1 | P2 | Avg. |
|---|---|---|---|
| All (C1 ~ C4, L1, L2) | 85.23 | 96.00 | 90.62 |
| **C3, C4, L1** | **87.92** | **96.00** | **91.96** |

Table 4. LSTM-HMM bases speech recognition results of speaker adaptation (word accuracy, %).

| | P1 | P2 | Avg. |
|---|---|---|---|
| LSTM + HMM | 86.58 | 96.67 | 91.63 |
| **Drop-out applied** | **89.33** | **97.33** | **93.63** |

200개의 뉴런을 갖고 3개의 은닉 층을 갖는 LSTM Cell 5개와 시그모이드 함수를 활성함수로 사용하였다. 모델 적응 과정에서 성능 향상을 위해 드롭 아웃 규제 기법을 적용하여 성능을 관찰하였다. 실험 결과 초기모델 80 %, 사전모델 70 %, 화자적응 50 %의 비율을 적용하였을 때 가장 높은 성능을 나타냈다. 결과적으로 DNN-HMM 대비 3.68 %의 인식률 향상을 보여 LSTM-HMM 구조가 완전 연결 형태의 DNN-HMM 구조보다 효과적임을 확인할 수 있었다. 이와 같은 결과는 시간에 따른 입력 데이터의 변화를 반영하는 순환 신경망 형태인 LSTM 구조가 VPI 환자 음성 인식에 효과적임을 증명하고, 모델 적응 단계에서 적용하는 드롭 아웃 규제 기법이 화자 모델 적응 성능을 높이는 데 도움이 됨을 입증한다.

## VI. 결 론

본 논문에서는 VPI 환자의 음성을 효과적으로 인식하기 위해 CNN, LSTM 구조의 음성 인식 시스템을 구축하고 모델 적응 기법을 적용한 모델과 기존 GMM-HMM 기반의 음성 인식 시스템 및 완전 연결 형태의 DNN-HMM 하이브리드 기반의 시스템과 성능을 비교하였다. 정상인 화자가 PBW452단어를 발화한 데이터를 이용하여 초기 모델을 학습하고 정상인 화자의 VPI 모의 음성을 이용하여 화자 적응의 사전 모델을 생성한 후에 VPI 환자들의 음성으로 추가 적응 훈련을 진행하였다. VPI 환자의 화자 적응 시에 CNN-HMM 기반 모델에서는 일부 층만 부분적으로 학습하였고, LSTM-HMM 기반 모델에서는 드롭 아웃 규제 기법 적용 후 성능을 관찰하여 DNN-HMM 시스템과 비교하였다.

VPI 환자의 화자 적응 시 CNN-HMM 기반의 모델의 경우 일부 층만을 적응 학습하여 91.96 %의 인식률을 보였고, LSTM-HMM 기반의 모델의 경우 모든 층을 적응 학습하고 드롭 아웃 규제를 적용하였을 때 93.63 %의 인식률을 나타내어 완전 연결형 DNN-HMM 기반의 모델보다 평균 3.68 %의 인식 성능이 향상하였다. 본 논문에서 제안하는 LSTM-HMM 기반의 하이브리드 음성 인식 기법이 많은 데이터를 확보하기 어려운 VPI 환자 음성에 대해 보다 향상된 인식률의 음성 인식 시스템을 구축하는 데 효과적임을 입증한다.

향후 연구에서는 CNN, LSTM 등 심층 신경망 단독 시스템에서 VPI 환자 음성 인식 성능을 평가하고 향상 방안을 연구할 계획이다.

## 감사의 글

이 논문은 2016년도 정부(미래창조과학부)의 재원으로 한국연구재단 이공학 개인기초 연구지원사업의 지원을 받아 수행된 연구임(NRF-2016R1D1A2B03935008).

## References

1. S. G. Fletcher, “Theory and instrumentation for quantitative measurement of nasality,” J. Cleft Palate. **7**, 601-609 (1970).
2. J. -E. Lee, W. -E. Kim, K. H. Kim, M. -W. Sung, and T. -K. Kwon,“Research on construction of the Korean speech corpus in patient with velopharyngeal insufficiency” (in Korean), JKORL. **55**, 498-507 (2012).
3. M. Y. Sung, H. Kim, T. -K. Kwon, and M. -W. Sung, “Analysis on vowel and consonants sounds of patient’s speech with velopharyngeal insufficiency (VPI) and simulated speech” (in Korean), JKIICE. **18**, 1740-1748 (2014).
4. M. Y. Sung, T. -K. Kwon, M. -W. Sung, and W. Kim, “Effective recognition of velopharyngeal insufficiency (VPI) patient's speech using simulated speech model” (in Korean), JKIICE. **19**, 1243-1250 (2015).
5. K. Yoon and W. Kim, “Effective recognition of velopharyngeal insufficiency (VPI) patient's speech using DNN-HMM-based system” (in Korean), JKIICE. **23**, 33-38 (2019).
6. *HTK Speech Recognition Toolkit*, http://htk.eng.cam.ac.uk/, (Last viewed March 11, 2015).
7. ETSI ES 201 108, *Standard Document*, v1.1.2.(2000-04)., 2000.

8. J. L. Gauvain and C. H. Lee, “Maximum a posteriori estimation for multivariate Gaussian mixture observations of markov chains,” IEEE Trans. on Speech and Audio Proc. **2**, 291-298 (1994).
9. C. J. Leggetter and P. C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density HMMs,” Computer Speech and Language, **9**, 171-185 (1995).
10. J. -T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, “Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers,” Proc. IEEE ICASSP. 7304-7308 (2013).
11. W. Hu, Y. Qian, and F. K. Soong, “A DNN-based acoustic modeling of tonal language and its application to Mandarin pronunciation training,” Proc. IEEE ICASSP. 3206-3210 (2014).
12. S. Park, Y. Jeong, and H. S. Kim, “Multiresolution CNN for reverberant speech recognition,” Proc. 20th Conf. O-COCOSDA. 1-4 (2017).
13. A. Senior, H. Sak, and I. Shafran, “Context dependent phone models for LSTM RNN acoustic modeling,” Proc. IEEE ICASSP. 4585-4589 (2015).
14. S. Hochreiter and J. Schmichuber, “Long short-term memory,” Neural Computation, **9**, 1735-1780 (1997).
15. S. J. Rennie, V. Goel, and S. Thomas, “Annealed dropout training of deep networks,” Proc. IEEE SLT. 159-164 (2014).

## 저자 약력

▸ 김 민 석 (Min-seok Kim)

2014년 3월 ~ 현재 : 인천대학교 컴퓨터 공학부 학사과정

▸ 정 재 희 (Jae-hee Jung)

2017년 3월 ~ 현재 : 인천대학교 컴퓨터 공학부 학사과정

▸ 정 보 경 (Bo-kyung Jung)

2017년 3월 ~ 현재 : 인천대학교 컴퓨터 공학부 학사과정

▸ 윤 기 무 (Ki-mu Yoon)

2018년 2월 : 인천대학교 수학과 학사
2018년 3월 ~ 현재 : 인천대학교 컴퓨터공학부 석사과정

▸ 배 아 라 (Ara Bae)

2019년 2월 : 인천대학교 컴퓨터공학부 학사
2019년 3월 ~ 현재 : 인천대학교 컴퓨터공학부 석사과정

▸ 김 우 일 (Wooil Kim)

1996년 2월, 1998년 8월, 2003년 8월 : 고려대학교 전자공학과 학/석/박사
2012년 8월 ~ 현재 : 인천대학교 컴퓨터공학부 조교수, 부교수