ISSN 2005-8063
2017. 6. 30.
Vol.9 No.2
pp. 95-101

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

http://dx.doi.org/10.13064/KSSS.2017.9.2.095

원어민 및 외국인 화자의 음성인식을 위한 심층 신경망 기반 음향모델링*

DNN-based acoustic modeling for speech recognition of native and foreign speakers

강 병 옥 · 권 오 욱**

Kang, Byung Ok · Kwon, Oh-Wook

Abstract

This paper proposes a new method to train Deep Neural Network (DNN)-based acoustic models for speech recognition of native and foreign speakers. The proposed method consists of determining multi-set state clusters with various acoustic properties, training a DNN-based acoustic model, and recognizing speech based on the model. In the proposed method, hidden nodes of DNN are shared, but output nodes are separated to accommodate different acoustic properties for native and foreign speech. In an English speech recognition task for speakers of Korean and English respectively, the proposed method is shown to slightly improve recognition accuracy compared to the conventional multi-condition training method.

Keywords: automatic speech recognition, Deep Neural Network (DNN), acoustic model

1. 서론

심층 신경망(Deep Neural Network; DNN)은 음성 및 이미지 인식 등의 전통적인 패턴인식 응용 분야에 이미 기본적으로 채택되어 적용되고 있고, CNN (Convolutional Neural Network) 및 LSTM-RNN (Long Short-Term Memory − Recurrent Neural Network) 등으로 확장되어 대화처리 및 자동번역 등으로 적용 범위를 확장해 가고 있다[1],[2]. 현재 상용화된 대부분의 음성인식 시스템은 기존에 사용되던 가우시안 혼합 모델-은닉 마코프 모델(Gaussian Mixture Model − Hidden Markov Model; GMM-HMM) 기반의 음향모델을 DNN, CNN 및 LSTM-RNN 기반의 음향모델로 대체하여 사용하고 있으며, DNN 기반 음향모델(acoustic model; AM)의 성능 개선 연구를 꾸준히 진행하고 있다. 최근 CTC (Connectionist Temporal Classification) 기반의 end-to-end 음성인식을 통해 HMM을 완전히 DNN으로 대체하기 위한 방법도 활발히 연구되고 있다. 하지만, 현재 서비스되고 있는 대부분의 음성인식 시스템은 심층 신경망-은닉 마코프 모델(Deep Neural Network − Hidden Markov Model; DNN-HMM) 기반의 음향모델을 채택하여 적용하고 있다.

현재의 DNN-HMM 기반 음향모델의 학습과정에서는, GMM-HMM 음향모델 학습을 통해 DNN 모델의 출력 노드 혹은 타겟에 해당하는 HMM 상태를 결정하고, 훈련 음성 데이터의 상태정렬 정보를 추출하는 과정이 선행된다. 즉, DNN 학습 과정은 HMM 학습 결과로 이미 결정된 훈련 음성데이터의 상태 정렬 정보를 받아서, 훈련 음성을 가장 잘 표현하는 특징 추출과 변별력 있는 모델 파라미터를 얻는 과정이라고 볼 수 있

* 본 연구는 미래창조과학부 및 정보통신기술진흥센터의 정보통신·방송 연구개발사업[지원번호: RO126-15-1117, 언어학습을 위한 자유발화형 음성대화처리 원천기술개발]과 2017년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업[No. 2015R1D1A3A01020817]의 일환으로 수행되었음

** 충북대학교, owkwon@cbnu.ac.kr, 교신저자

Received 5 May 2017; Revised 1 June 2017; Accepted 13 June 2017

다. 이때 상태정렬 정보를 DNN 학습과정에 포함시켜 주기적인 상태 정렬을 수행한 후 반복학습을 하는 방법도 있지만, 이미 앞 단에서 결정한 출력 노드, 즉 상태는 변경이 되지 않는다.

한편, 대어휘 음성인식을 위한 음향모델을 구성하는 HMM 상태는 일반적으로 음소의 문맥정보를 바탕으로 언어학자가 설계한 질의어를 통해 최대 로그우도 기준으로 분기되는 결정트리(decision tree) 기반 방식[3]을 통해 말단 노드의 HMM 상태가 공유되는 형태로 결정된다. 하지만, 원어민 및 외국인 화자를 인식 대상으로 하는 음향모델(예를 들어, 영어/한국어/중국어를 모국어로 사용하는 화자를 대상으로 하는 영어 음성인식기의 경우)과 같이, 서로 다른 음향적 특성을 갖는 대규모 훈련 음성 데이터를 대상으로 하나의 결정트리를 통해 상태를 결정하는 방식은 비효율적이다[4]. 그 이유는 동일한 음소 문맥에 대해서도 비원어민과 원어민 음성의 음향적 특성은 차이가 있고, 비원어민 음성 사이에서도 각각의 모국어에 따라 다른 음향적 특성을 가지고 있기 때문이다. 따라서 이러한 기존의 전체 훈련용 음성데이터를 입력으로 하여 하나의 결정트리로 상태를 결정하는 방법의 문제점을 개선하기 위한 방법들이 제안된 바 있다[4],[5],[6].

본 논문에서는 원어민과 외국인의 발화로 구성된 훈련용 음성데이터와 같이 서로 다른 음향적 통계 특성을 갖는 다중 집합 훈련용 음성데이터를 대상으로 DNN 기반 음향모델을 학습하기 위해, 먼저 다중 집합 상태 클러스터를 결정하고, 은닉층 파라미터를 공유하면서도 다중 집합 상태 클러스터를 수용하는 출력 노드들로 구성된 출력층을 갖는 DNN 기반 음향모델 구조를 제안한다. 여기서 다중 집합 상태 클러스터는 서로 다른 음향적 특성을 갖는 개별 집합의 훈련용 음성데이터를 대상으로 각각의 최적 상태 집합을 생성한 후에, 전체 집합을 대상으로 제안하는 방법에 따라 유사 상태를 병합하여 재구성한 상태 클러스터를 의미한다. 인식단계에서는 미리 정해진 특정 모국어를 사용하는 화자의 입력 음성에 대해 해당 집합 상태 클러스터로 구성된 더 적은 수의 출력 노드를 갖는 최적화된 DNN 음향모델을 기반으로 인식을 수행함으로써, 개선된 인식 성능을 기대할 수 있다. 본 논문에서 제안한 방법을 한국인과 원어민이 발성한 훈련용 영어 음성 데이터로 구성한 음향모델 학습에 적용하여 평가한 결과 원어민 발화 영어 음성에 대한 음성인식 성능을 유지하면서, 한국인 발화 영어 음성에 대해 개선된 음성인식 성능을 얻을 수 있었다.

본 논문의 구성은 다음과 같다. 2 장에서는 본 논문에서 제안하는 원어민 및 외국인의 음성인식을 위한 다중 집합 상태 클러스터를 기반으로 한 DNN 음향모델에 대해 설명한다. 3 장에서는 본 논문에서 제안한 방법을 평가하기 위해 한국인 및 원어민 대상 영어 음성인식을 위한 DNN 기반 음향모델을 대상으로 한 실험환경과 성능평가 결과에 대해 설명한다. 4 장에서는 결론과 향후 연구방향에 대해서 기술한다.

2. 다중 집합 상태 클러스터를 기반으로 한 DNN 음향모델

훈련용 음성데이터가 서로 다른 음향적 특성을 갖는 경우, 예를 들어 중국인/한국인과 같은 외국인과 원어민의 영어 발화로 구성되어 있다고 가정할 때, 기존의 다중 조건 기반 음향모델 학습 및 인식 방법은 <그림 1>과 같이 수행된다.

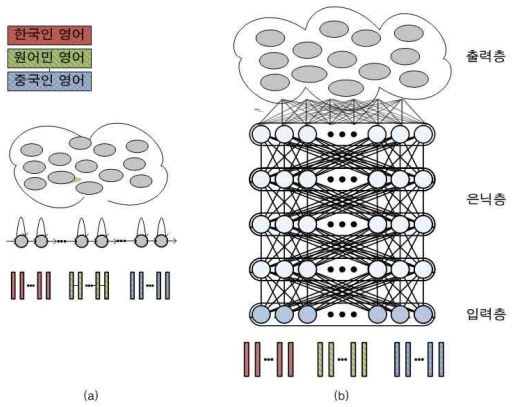

그림 1. 기존의 다중 조건 기반 음향모델링 개념도
(a) 단일 결정트리에 의한 상태 결정 단계,
(b) 음향모델 훈련 및 인식 단계
Figure 1. Conceptual diagram of the conventional multi-condition based acoustic modeling
(a) State decision step by single decision tree,
(b) Training and decoding step

기존 다중 조건 기반 음향모델 학습 방법은 <그림 1>에서와 같이 다양한 음향 특성을 가진 전체 훈련용 음성데이터를 대상으로 하나의 결정 트리에 의해 적절한 음향모델 단위를 결정하고, 이를 대상으로 음향모델 학습 및 인식을 수행한다.

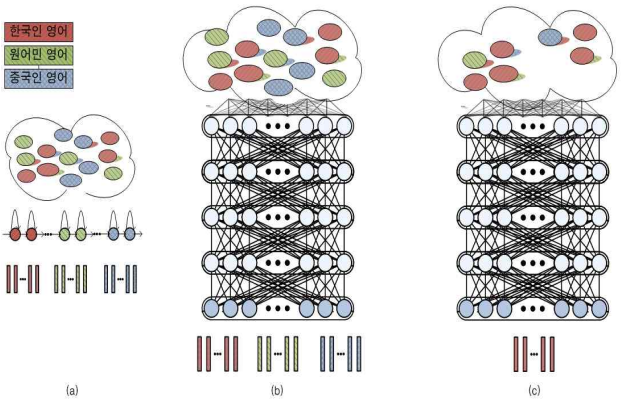

그림 2. 다중 집합 음향 공간 모델링을 위한 심층신경망 기반 음향모델링 개념도
(a) 다중 집합 상태 클러스터링 단계, (b) 다중 집합 상태 클러스터 대상 훈련 단계, (c) 다중 집합 상태 클러스터 대상 인식 단계 (선택된 화자에 해당하는 상태 집합 대상 인식)
Figure 2. Conceptual diagram of DNN-based acoustic modeling for multi-set acoustic space
(a) Multi-set state clustering step, (b) Training step based on multi-set state cluster, (c) Decoding step based on multi-set state cluster (decoding for selected state classes of input speaker)

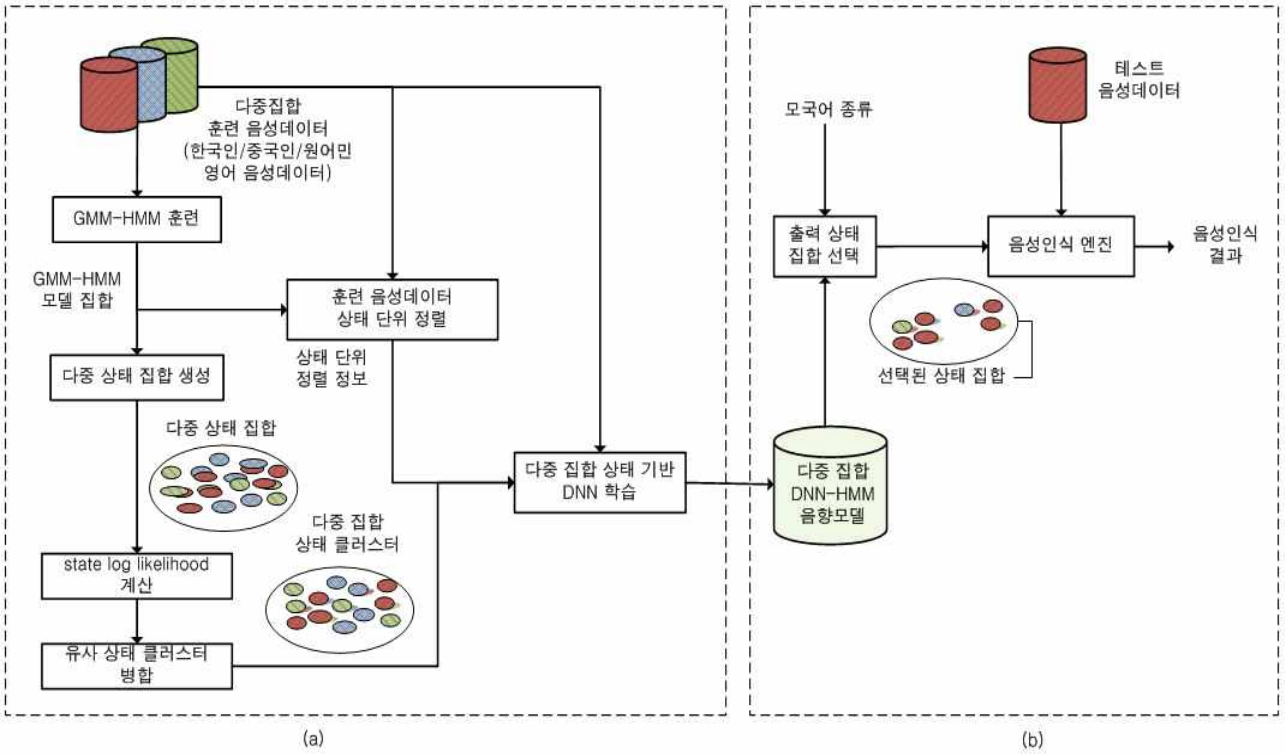

그림 3. 다중 집합 음향 공간 모델링을 위한 심층신경망 기반 음향모델링 블록도
(a) 다중 집합 상태 클러스터링 및 훈련 단계, (b) 인식 단계 (선택된 화자에 해당하는 상태 집합 대상 인식)
Figure 3. Block diagram of DNN-based acoustic modeling for multi-set acoustic space
(a) Multi-set state clustering and training step, (b) Decoding step (decoding for selected state classes of input speaker)

동일한 구성의 훈련용 음성데이터에 대해 본 논문에서 제안하는 다중 집합 상태 클러스터를 기반으로 한 DNN 음향모델 학습 및 인식방법의 개념도는 <그림 2> 와 같다.

첫 번째로 다중 집합 상태 클러스터 생성단계(a)에서는 개별 모국어 화자로 구성된 훈련용 음성데이터를 대상으로 각각의 최적 상태 집합을 생성한 후에 전체 상태 집합을 대상으로 클러스터링을 수행한다. 이때 비슷한 음향공간을 공유하는 상태들은 공통의 상태로 결합되는데, 이와 같이 공유되는 상태를 서로 다른 색상의 그림자를 통해 표현하였다. 개별 색상으로 표현되는 공유되지 않는 상태는 독립적인 음향공간을 차지한다. 예를 들어 한국어 모국어 화자의 영어 음향모델 상태는 단독 상태 공간을 차지하거나 다른 모국어 화자의 영어 음향모델 상태공간을 공유한다. 훈련단계(b)에서는 생성단계에서 생성된 다중 집합 상태 클러스터를 출력 노드로 하여 DNN 기반 음향모델 훈련을 수행한다. 인식단계(c)에서는 개별 모국어 화자 집합의 입력 음성에 대해서는 해당 상태 클러스터로 구성된 출력 노드를 대상으로 음성인식이 수행된다.

<그림 3>은 <그림 2>를 구체화하기 위해, 다중 집합 상태 클러스터를 기반으로 한 DNN 음향모델 학습 및 인식방법에 대한 순서도를 보여준다. 각 단계별로 상세한 설명은 다음과 같다.

2.1. 다중 집합 상태 클러스터 생성

앞 절과 마찬가지로 다중집합 훈련데이터는 영어/한국어/중국어와 같이 서로 다른 모국어를 갖는 화자의 영어 발화 음성데이터를 예시로 한다. 각각의 훈련 음성데이터를 대상으로 음향모델 훈련을 통해 서로 다른 최적의 결정트리와 상태집합을 갖는 GMM-HMM 모델집합을 생성한다. 각 상태집합의 상태들을 모아서 구성한 상태 클러스터를 구성하고, 전체 상태 집합의 개별 상태들 중에 비슷한 음향공간을 차지하는 상태들을 대상으로 클러스터링 과정을 거쳐 다중 집합 상태 클러스터를 생성한다[5],[6]. 클러스터링 과정은 두 가지 기준에 의해 수행된다.

첫 번째 기준은 다음과 같다.

$$\Delta L_{merged} = L(state_1) + L(state_2) - L(state_{merged}) \quad (1)$$

$$\Delta L_{merged} \leq Threshold$$

위 식에서 $L(state_1)$ 과 $L(state_2)$ 는 다중 상태 집합에서 선택된 임의의 두 상태의 로그우도이고, $L(state_{merged})$ 는 두 상태가 결합되었다고 가정된 상태에서 계산된 로그우도이다. 두 상태가 결합되었을 때의 로그 우도의 차 $\Delta L(state_{merged})$가 $Threshold$ 보다 적을 경우 두 상태는 훈련데이터의 관점에서 서로 비슷한 음향 공간을 차지한다고 볼 수 있으므로 병합이 가능해진다. 식 (1)에서 상태 및 상태 집합의 로그우도 $L(S)$는 식 (2)를 통해 구해진다[3].

$$L(S) = \sum_{s\in S}\sum_{f\in F_s} \log\left(P(x_f;\mu_s,\Sigma_s)\right)\gamma_s(x_f) \quad (2)$$
$$= -\frac{1}{2}\left(\log\left[(2\pi)^n|\Sigma_S|\right]+n\right)\sum_{s\in S}\sum_{f\in F_s}\gamma_s(x_f)$$

식 (2)에서 $L(S)$ 는 해당 상태의 로그우도, S 는 클러스터링된 상태 집합, F_s는 각 상태 s를 구성하는 프레임의 수, f 는 각 특징벡터 프레임, n 은 특징벡터의 차원, $\gamma_s(x_f)$ 는 상태 s의 관측된 특징벡터 x_f의 사후관측확률을 의미한다. $P(x_f;\mu_s,\Sigma_s)$는 상태관측확률을 의미하고, 가우시안 분포의 경우 두 번째 식으로 계산이 가능하다. 여기서 μ_s 는 각 상태 s의 평균, Σ_s 는 각 상태 s의 공분산을 의미한다. 식 (2) 두 번째 줄의 Σ_S는 클러스터링된 상태 집합 S의 공분산이고, 해당 상태 집합 클러스터를 구성하는 개별 상태의 평균과 공분산으로 계산된다.

두 번째 기준은 다음과 같다.

$$\{T_{set1}|T_{set1} \text{ is the logical triphone sharing } state_1\} \quad (3)$$
$$= \{T_{set2}|T_{set2} \text{ is the logical triphone sharing } state_2\}$$

다중 상태 집합을 구성하는 상태들은 각각을 생성한 결정트리에 의해 결정된, 공유하는 음향공간을 구성하는 논리적인 트라이폰(logical triphone) 집합을 가지고 있다. 서로 다른 결정 트리에 의해 생성된 상태일 경우에도, 식 (3)과 같이 다중 상태 집합에서 선택된 임의의 두 상태 $state_1$과 $state_2$를 공유하는 논리적인 트라이폰의 집합이 같을 경우 두 상태는 서로 비슷한 음향공간을 차지한다고 볼 수 있으므로 병합이 가능하다.

유사 상태 병합으로 최종적으로 생성된 상태들은 <그림 3>의 훈련단계에서 다중 집합 상태 클러스터와 같이 도식적으로 표현이 가능하다. 즉, 단독 상태 공간을 구성하는 경우가 있을 수 있고, 서로 다른 집합의 상태와 결합되어 상태공간을 공유하는 상태가 있을 수 있다.

2.2. DNN 음향모델 훈련 및 음성인식

각 집합의 훈련 음성데이터에서 얻어진 상태 정렬 정보로부터 얻은 훈련 음성데이터/레이블을 입력 데이터로 하고, 다중 집합 상태 클러스터로부터 얻어진 상태들을 출력 노드로 하여, 다중 집합 상태 기반 DNN 학습을 수행한다.

본 논문에서 제안하는 다중 집합 상태 클러스터 기반 DNN 음향모델 훈련 방법은 특정 집합 훈련 음성데이터, 예를 들어 한국어 모국어 화자의 훈련용 영어 음성 데이터 집합에 대해서는 해당 음성 프레임을 입력으로 하고, 한국어 모국어 화자의 영어 음향모델 상태집합을 출력 노드로 하여 학습을 수행한다. 같은 방법으로 다른 모국어 화자의 훈련용 음성 데이터에 대해서는 해당 음성 프레임을 입력으로 하고, 해당 모국어 화자의 영어 음향모델 상태 집합을 출력 노드로 하여 학습이 진행된다. 이 과정에서는 입력 노드 집합과 관계가 없는 다른 상태 클러스터로의 출력 노드 연결을 끊고, 해당 집합 상태 클러스터로의 출력 노드 연결이 학습됨으로써 더 최적화된 형태의 학습을 기대할 수 있다.

인식 단계에서는 DNN 훈련단계에서 학습하여 얻어진 다중 집합 상태 클러스터를 출력 노드로 갖는 DNN-HMM 구조의 음향모델을 사용하여 인식을 수행하게 된다. 음성인식기는 발화자의 모국어를 알고 있다고 가정하면, 인식 단계에서 전체 다중 집합 상태 클러스터 중에 특정 모국어에 해당하는 음향모델 상태 집합만을 출력 노드로 하는 음성인식이 가능하다. 이러한 과정은 <그림 3>의 인식 단계에서 도식적으로 표현된다.

3. 실험 및 결과

본 논문에서 제안하는 다중 집합 상태 클러스터를 기반으로 한 DNN 음향모델 학습 및 인식방법을 평가하기 위해, 한국인과 원어민이 발성한 훈련용 영어 음성 데이터로 구성한 음향모델 학습 및 성능평가에 적용하였다.

음향모델 훈련에 사용된 원어민 음성 데이터는 Linguistic Data Consortium (LDC), 음성정보기술산업지원센터(SiTEC), 한국전자통신연구원(ETRI) 등에서 수집된 전체 426 시간 분량의 미국 원어민 발성 영어 음성 데이터를 사용하였다. 음향모델 훈련에 사용된 한국인 발성 음성 데이터는 한국전자통신연구원(ETRI)에서 수집한 전체 382 시간 분량의 한국인 발성 영어 음성데이터를 사용하였다. 음향모델 훈련에 사용된 원어민/한국인 발성 영어 음성데이터는 16 kHz 샘플링 주파수의 마이크 음성데이터로서 다양한 채널, 도메인, 화자로부터 수집된 단어 단위 또는 문장 단위로 발성한 음성데이터로 구성된다.

GMM-HMM 음향모델 훈련에 사용된 음성데이터 특징벡터는 감마톤 필터뱅크 기반의 켑스트럼 특징을 사용하고[7],[8], DNN 음향모델 훈련을 위해서는 40 차 로그 mel-scale 필터뱅크 [9] 출력을 사용하였다. DNN 음향모델을 구성하는 은닉층은 5 개 층을 사용하였고, 은닉층의 노드 수는 2,048 개를 기본으로 하고 비교를 위해 1,024 개에 대해서도 실험을 하였다. 훈련 음성데이터에 대해서는 특징벡터에 사용된 로그 mel-scale 필터뱅크 출력의 차원인 40 차에 대해 현재프레임과 앞 뒤 7 개 프레임으로 구성된 15 개 프레임의 문맥 창(context window)이 사용된 총 600 차의 입력데이터를 구성하였고, 입력 데이터 정규화를 위해 차원 감소 없는 LDA (Linear Discriminant Analysis)를 수행한 결과를 DNN 음향모델의 입력층(input layer)으로 사용하였다.

본 논문에서 제안하는 다중 집합 상태 클러스터를 기반으로 한 DNN 음향모델을 평가하기 위해, 기존의 방법에 따라 한국인 발화와 원어민 발화의 영어 음성데이터 전체를 대상으로 단일한 결정트리를 생성하여 DNN-HMM 기반 음향모델을 훈련하는 다중 조건(multi-condition) 기반 음향모델을 기본 모델(baseline model)로 하여 비교 평가를 수행하였다. 다중 조건 기반 DNN 음향모델은 선행되는 GMM-HMM 학습과정을 통해 11,485 개 상태를 생성하고, 전체 훈련용 음성데이터를 대상으로 상태 정렬을 수행하였다. 이렇게 결정된 음성프레임/레이블

쌍으로 구성된 입력데이터와 11,485 개의 노드로 구성된 출력층(output layer) 대상으로 DNN 학습을 수행하였다.

다중 집합 상태 클러스터를 기반으로 한 DNN 음향모델을 생성하기 위해, 먼저 한국인 발성 영어와 원어민 발성 영어로 구성된 각 집합의 훈련데이터를 대상으로 GMM-HMM 음향모델을 생성하였다. 원어민 발화 영어 음성데이터를 대상으로 결정트리를 생성한 후 그 결과 생성된 9,326 개의 상태로 훈련된 GMM-HMM 기반 음향모델을 훈련하고, 이를 기반으로 상태정렬을 수행하였다. 또한 한국인 발화 영어 음성데이터를 대상으로 동일한 기준으로 9,033 개의 상태로 구성되는 GMM-HMM 기반 음향모델을 훈련하고, 이를 기반으로 상태정렬을 수행하였다. 전체 상태 집합의 상태들을 대상으로 제안된 방법을 통해 클러스터링 과정을 거쳐 다중 집합 상태 클러스터를 생성하고, 이를 기반으로 각각의 훈련 음성데이터의 상태정렬 정보는 유지하되, 음성프레임/레이블 쌍에 대해 클러스터링 된 상태정보에 따라 레이블을 갱신하였다. 갱신된 전체 음성프레임/레이블 쌍으로 구성된 입력데이터와 다중 집합 상태 클러스터링 정보가 반영된 출력층을 대상으로 DNN 학습을 수행하였다.

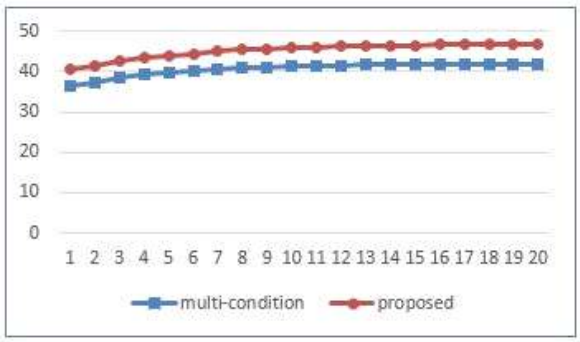

그림 4. 한국인 발화 영어 데이터의 프레임 인식률
Figure 4. Frame accuracy of Korean-speaking English DB

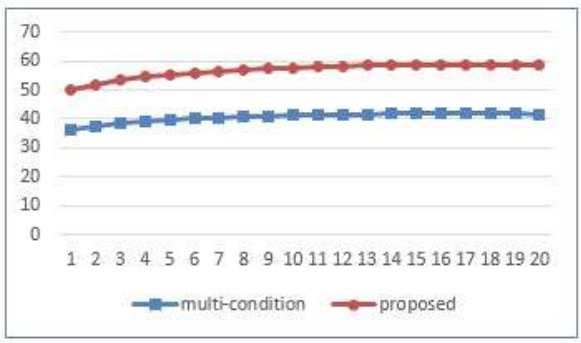

그림 5. 원어민 발화 영어 데이터의 프레임 인식률
Figure 5. Frame accuracy of native-speaking English DB

<그림 4>와 <그림 5>는 비교를 위해 훈련한 다중조건 DNN 기반 음향모델과 본 논문에서 제안하는 다중 집합 상태 클러스터를 기반으로 한 DNN 기반 음향모델의 학습과정에서 평가 데이터의 일부를 대상으로 각 반복 학습 단계에서 측정한 프레임 인식률 추이이다. 학습과정에서의 프레임 인식률은 각 입력 프레임의 classification 성능을 측정하는 지표로서, DNN-HMM 기반 음향모델을 적용한 음성인식의 경우와 같이 발음 사전 내의 음소 간 천이 및 HMM 상태 천이 확률이 적용되지 않는다. 따라서 프레임 인식률의 개선 정도에 비례하여 최종 목적인 인식 성능의 개선으로 반영되는 것을 보장하지는 않지만, 본 논문에서 제안하는 음향모델 학습 방법의 1 차 검증 방법으로 의미가 있다고 할 수 있다. <그림 4>는 한국인 발화 영어 데이터 대상 프레임 인식률이고, <그림 5>는 원어민 발화 영어 데이터 대상 프레임 인식률로서, 제안된 방법이 기존 다중 조건 기반 방법에 비해 일관되게 높은 프레임 인식률을 보이고 있다.

이는 동일한 조건의 hyper-parameter로 구성한 DNN 음향모델의 입력층과 은닉층에 대해, 기존의 다중 조건 기반 DNN의 출력층에 비해 제안된 방법이 더 최적화된 상태로 구성된 출력층으로 구성되고, 그 결과 개선된 프레임 인식률을 갖는 것으로 해석된다.

본 논문에서 제안된 방법으로 생성한 음향모델을 대상으로 한국인 발화 영어와 원어민 발화 영어 음성으로 평가셋을 구성하여 음성인식 비교 평가를 수행하였다. 음향모델 학습 및 음성인식에 사용되는 발음사전을 위해서는 CMU-DICT[10]에서 제안한 39 phone set을 사용하였다. 평가를 위해 사용된 언어모델은 55k 어휘에 대해 2.5M 개 bigram과 2.7M개 trigram으로 구성된 언어모델을 사용하였다. 한국인 발화 영어 평가데이터는 ETRI에서 연구개발한 음성인식 및 대화처리 기술이 적용된 한국인 대상 영어 학습 응용 프로그램인 GinieTutor [11]의 평가를 위해 수집한 음성데이터로 구성되어 있다. 대화를 통한 영어 학습 목적의 특성상 일정 수준 이상의 인식성능이 나와야 학습이 원활이 진행된다. 따라서 해당 도메인과 언어학습 콘텐츠에 최적화된 어휘 및 언어모델을 평가에 사용하였다. 음성인식 엔진은 ETRI에서 개발한 FST (finite state transducer) 기반의 대어휘 음성인식 시스템을 사용하였다[12]. 한국인 발화 영어를 위한 평가데이터는 각각 1,500 개와 1,200 개 발화로 구성된 두 개의 평가셋을 사용하였고, 평가 결과는 <표 1>과 같다.

표 1. 한국인 발화 영어 평가데이터 대상 음성인식률 비교
Table 1. Comparison of speech recognition accuracy of Korean-speaking English test DB

AM	# hidden nodes	KorEng Set1	KorEng Set2
Multi-condition	2,048	94.20	95.91
Proposed	1,024	94.24	96.03
	2,048	95.04	96.27

한국인 발화 영어로 구성된 평가셋을 대상으로 본 논문에서 제안된 방법에 따른 음향모델을 성능 평가한 결과 기존의 다중 조건 음향모델에 비해 개선된 인식 성능을 보임을 확인할 수 있다. 특히, DNN 모델의 hyper-parameter 중 은닉층의 노드 수를 기존의 2,048 개에서 1,024 로 줄여도 기존 다중 조건 음향모델과 비슷한 성능을 유지함을 볼 수 있다. 이는 서로 다른 음향적 특성을 갖는 전체 원어민/한국어 발화 영어 훈련데이터에 대해

특징 추출 기능이 강한 DNN 하부의 은닉층을 공유하여 더 다양한 특성이 반영된 강건한 모델을 제공하는 동시에, 서로 다른 상태간의 변별력이 강조되는 출력층에 대해서는 음향적 특성이 비교적 균일한 한국인 발화 영어 훈련데이터로 생성된 결정트리와 이를 통해 결정된 상태로 구성함으로써 비슷한 특성의 평가데이터에 대해서 더 강한 변별력을 제공한 효과로 해석된다.

표 2. 원어민 발화 영어 평가데이터 대상 음성인식률 비교
Table 2. Comparison of Speech recognition accuracy of native-speaking English test DB

AM	# hidden nodes	NatEng Set
Multi-condition	2,048	78.59
Proposed	1,024	78.44
	2,048	79.04

원어민 발화 영어를 위한 평가셋은 4,878 개 발화로 구성된 WSJ1 음성데이터[13]의 평가셋을 사용하였고, 평가 결과는 <표 2>와 같다. 평가에 사용된 언어모델과 어휘 셋은 영어 학습용으로 학습한 언어모델과 어휘 셋으로서, 영어 학습용 콘텐츠를 위한 텍스트를 포함하면서도, 대화를 통한 영어 학습을 위해 사용자의 다양한 패턴의 자유 발화를 인식할 수 있도록 WSJ 코퍼스를 포함한 다양한 영어 텍스트 코퍼스를 반영한 후 생성한 언어모델과 어휘 셋을 사용하였다. 동일한 조건의 언어모델과 어휘 셋을 사용하여 인식성능 비교를 수행한 결과 기존의 다중 조건 음향모델과 비슷한 수준을 유지하거나 미세하지만 약간의 성능 개선이 있다고 해석될 수 있다.

4. 결론

본 논문에서는 원어민 및 외국인 화자를 위한 음성인식을 위한 음향 모델 생성 방법으로 다중 집합 상태 클러스터 기반 DNN 기반 음향모델 훈련/인식 방법을 제안한다. 즉, 서로 다른 음향적 특성을 갖는 다중 집합 음성데이터에 대해 서로 비슷한 음향공간을 차지하는 상태를 결합한 다중 집합 상태 클러스터를 생성하고, 특징 추출 기능을 수행하는 DNN 하부 은닉층을 공유하되 출력층에 대해서는 변별력이 있는 균일한 집합의 상태를 갖는 DNN 구조와 학습/인식 방법을 제안하였다. 이를 한국인/원어민 발화 영어 음성인식에 적용하여 평가한 결과 기존의 다중 조건 기반 음향모델에 비해 개선된 인식성능을 얻을 수 있었다.

향후 연구로, 상태 결정 과정과 DNN 학습과정을 결합하여, 전체 DNN 학습과정 내에서 DNN 파라미터로 계산되는 기준에 따라 상태를 결정하고, 결정된 상태에 따라 DNN 파라미터 훈련을 수행하는 학습방법에 대한 연구를 진행할 예정이다.

참고문헌

[1] Abdel-Hamid, O., Mohamed, A., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(10), 1533-1545.

[2] Sak, H., Senior, A., & Beaufays, F. (2014). Long short-term recurrent neural network architectures for large scale acoustic modeling. *Proceedings of INTERSPEECH-2014* (pp. 338-342). 2014.

[3] Young, S. J., Odell, J. J., & Woodland, P. C. (1994). Tree-based state tying for high accuracy acoustic modelling. *Proceedings of ARPA Human Language Technology Workshop* (pp. 307-312). 1994.

[4] Chen, X., & Cheng, J. (2012). Acoustic modeling for native and non-native Mandarin speech recognition. *Proceedings of International Symposium on Chinese Spoken Language Processing*. 2012.

[5] Kang, B., Jung, H., & Kwon, O. (2013). Noise robust spontaneous speech recognition using multi-space GMM. *Proceedings of INTERNOISE-2013*. Innsbruck, Austria. September, 2013.

[6] Kang, B., & Kwon. O. (2016). Combining multiple acoustic models in GMM spaces for robust speech recognition. *IEICE Transactions on Information and Systems*, 99(3), 724-730.

[7] Lee, S., Kang, B., Chung, H., & Lee, Y. (2014). Intra- and inter-frame features for automatic speech recognition. *ETRI Journal*, 36(3), 514-517.

[8] Lee, S., Kang, B., Chung, H., & Park, J. (2015). A useful feature-engineering approach for a LVCSR System based on CD-DNN-HMM Algorithm. *Proceedings of the 2015 European Signal Processing Conference* (pp. 1436-1440). September, 2015.

[9] Mohamed, A. R., Hinton, G., & Penn, G. (2012). Understanding how deep belief networks perform acoustic modelling. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (pp. 4274-4276). 2012.

[10] Carnegie Mellon University. Carnegie Mellon Pronunciation Dictionary. Retrieved from http://www.speech.cs.cmu.edu/cgi-bin/cmudict on March 2, 2015.

[11] Kwon, O., Lee, K., Roh, Y., Huang, J., Choi, S., Kim, Y., Jeon, H.,Oh, Y., Lee, Y., Kang, B., Chung, E., Park, J., & Lee, Y. (2015). GenieTutor: A computer assisted second language learning system based on spoken language understanding. *Proceedings of the International Workshop on Spoken Dialog System (IWSDS 2015)*. Busan, South Korea. January, 2015.

[12] Chung, H., Park, J., Jeon, H., & Lee, Y. (2009). Fast speech recognition for voice destination entry in a car navigation system. *Proceedings of INTERSPEECH-2009* (pp. 975-978). Brighton, UK. 2009.

[13] Paul, D. B., & Baker, J. M. (1992). The design for the Wall

Street Journal-based CSR corpus. *Proceedings of ICSLP-1992* (pp. 899-902). October, 1992.

• **강병옥 (Kang, Byung Ok)**
한국전자통신연구원 음성지능연구그룹
대전광역시 유성구 가정로 218
Tel: 042-860-5684
Email: bokang@etri.re.kr
관심분야: 음성인식, 음향모델

• **권오욱 (Kwon, Oh-Wook)** 교신저자
충북대학교 전자공학부
충북 청주시 서원구 충대로1
Tel: 043-261-3374
Email: owkwon@cbnu.ac.kr
관심분야: 음성인식, 음성 및 오디오 신호처리, 패턴 인식