한국음향학회지 제38권 제5호 pp. 607~613 (2019)
The Journal of the Acoustical Society of Korea Vol.38, No.5 (2019)
https://doi.org/10.7776/ASK.2019.38.5.607

pISSN : 1225-4428
eISSN : 2287-3775

External knowledge를 사용한 LFMMI 기반 음향 모델링

LFMMI-based acoustic modeling by using external knowledge

박호성,[1] 강요셉,[1] 임민규,[1] 이동현,[1] 오준석,[1] 김지환[1†]

(Hosung Park,[1] Yoseb Kang,[1] Minkyu Lim,[1] Donghyun Lee,[1] Junseok Oh,[1] and Ji-Hwan Kim[1†])

[1]서강대학교 컴퓨터공학과

(Received July 10, 2019; accepted September 10, 2019)

초 록: 본 논문은 external knowledge를 사용한 lattice 없는 상호 정보 최대화(Lattice Free Maximum Mutual Information, LF-MMI) 기반 음향 모델링 방법을 제안한다. External knowledge란 음향 모델에서 사용하는 학습 데이터 이외의 문자열 데이터를 말한다. LF-MMI란 심층 신경망(Deep Neural Network, DNN) 학습의 최적화를 위한 목적 함수의 일종으로, 구별 학습에서 높은 성능을 보인다. LF-MMI에는 DNN의 사후 확률을 계산하기 위해 음소의 열을 사전 확률로 갖는다. 본 논문에서는 LF-MMI의 목적식의 사전 확률을 담당하는 음소 모델링에 external knowlege를 사용함으로써 과적합의 가능성을 낮추고, 음향 모델의 성능을 높이는 방법을 제안한다. External memory를 사용하여 사전 확률을 생성한 LF-MMI 모델을 사용했을 때 기존 LF-MMI와 비교하여 14 %의 상대적 성능 개선을 보였다.

핵심용어: 음성인식, 음향모델, lattice 없는 상호 정보 최대화, 음소 기반 언어 모델

ABSTRACT: This paper proposes LF-MMI (Lattice Free Maximum Mutual Information)-based acoustic modeling using external knowledge for speech recognition. Note that an external knowledge refers to text data other than training data used in acoustic model. LF-MMI, objective function for optimization of training DNN (Deep Neural Network), has high performances in discriminative training. In LF-MMI, a phoneme probability as prior probability is used for predicting posterior probability of the DNN-based acoustic model. We propose using external knowledges for training the prior probability model to improve acoustic model based on DNN. It is measured to relative improvement 14 % as compared with the conventional LF-MMI-based model.

Keywords: Speech recognition, Acoustic model, LF-MMI (Lattice Free Maximum Mutual Information), Phoneme-based language

PACS numbers: 43.72.Bs, 43.72.Ne

I. 서 론

음성인식은 사람의 말소리를 입력 받아 해당 소리에 해당하는 기호의 열을 결과로 출력하는 것을 말한다. 음성 인식의 문제 정의는 음성 신호의 열(sequence)이 입력으로 주어졌을 때, 모델에서 가장 높은 확률을 보이는 기호의 열(word sequence)을 출력하는 것이다. 이를 수식으로 나타내면 아래와 같다.

$$\arg_w \max P(W|O). \quad (1)$$

Eq. (1)은 음성인식의 문제 정의를 나타낸다. O는 관측열을 말하며, 입력되는 음성의 열을 나타낸다. W는 단어열을 말하며, 출력되는 단어의 열을 나타낸다. 하지만 조건부 확률인 관측열의 확률은 가능한 음성 발성에 대한 경우의 수가 너무 많아 실질적으로 무한대이므로 확률을 계산할 수 없다. 이 문제를 해결하기 위해서 아래 식으로 변환하여 사용한다.

†**Corresponding author:** Ji-Hwan Kim (kimjihwan@sogang.ac.kr)
Department of Computer Science and Engineering, Sogang University, 35 Baekbum-ro, Mapo-gu, Seoul 04107, Republic of Korea
(Tel: 82-2-705-8924)

$$\arg_w \max \frac{P(O|W)P(W)}{P(O)}. \quad (2)$$

Eq. (2)는 Bayesian rule에 따라 Eq. (1)을 변형한 결과이다. Eq. (2)를 통해 음성 인식 문제의 사전 확률과 사후 확률을 구분할 수 있게 된다. *P(O|W)*의 확률을 계산할 수 있는 모델을 음향 모델이라고 하며, *P(W)*를 계산하는 모델을 언어 모델이라고 한다. 분모에 있는 *P(O)*의 경우 발생할 수 있는 경우의 수가 무한대이기 때문에 실질적으로 모델링할 수 없으나, 음성 인식의 문제는 정확한 확률을 출력하는 것이 아닌 가장 높은 확률을 가진 단어열을 출력하는 문제이다. 따라서, 가능한 *O*의 발생 확률이 모두 같다고 가정한다면 음성 인식 문제를 해결하는 데 있어 생략 가능하다. Eq. (2)에 따라, 음성 인식의 문제를 해결하기 위해 구현해야 하는 요소는 음향 모델, 언어 모델이 있으며, 이 두 가지 모델에서 얻어진 확률을 통해 가장 높은 확률의 단어열을 실제로 출력하는 디코딩 네트워크가 있다. 추가적으로, *W*의 인식 단위를 결정하는 단어 사전 또한 필요로 한다.

음성 인식에서 음향 모델이란, 모델이 주어졌을 때 입력된 발화의 생성 확률을 구하는 것을 말한다. 음향 모델링을 위해 가장 널리 사용되는 모델은 은닉 마코프 모델(Hidden Markov Model, HMM)이다. 이는 Markov chain을 기반으로 한 sequence modeling 방법으로, 음성인식 뿐 아니라 열을 다루는 문제의 해결법으로 널리 사용되고 있다.[1-3]

HMM을 통해 해결할 수 있는 문제는 인식, 강제 정렬, 학습이다. 인식은 모델이 주어졌을 때, 입력받은 관측열에 대한 확률을 계산하여 확률이 최대가 되는 모델을 선택하는 과정이다. 이 과정에서 HMM 생성 확률을 계산하기 위해 forward algorithm이 사용된다. 강제 정렬은 모델 학습을 하기 위한 전처리 과정으로, 전체 학습 자료에서 특정 단어를 발성하는 위치를 파악하여 모델 별 학습에 필요한 자료를 자동으로 추출한다. 이를 통해, 단어열로 주어진 데이터에 대해 인식 단위별 학습 자료를 생성할 수 있다. 강제 정렬을 수행하기 위해 viterbi algorithm이 사용된다. 학습은 자료가 주어졌을 때, 해당 자료의 확률이 최대가 되도록 모델 매개변수를 갱신하는 과정이다.

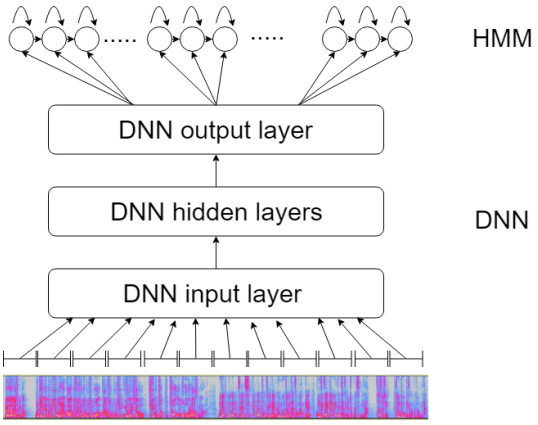

Fig. 1. DNN-HMM architecture in speech recognition.

주어진 학습 자료에 대하여 인식 과정을 수행할 수 있도록, 확률이 최대가 될 때 까지 HMM 매개변수를 갱신한다. 학습 과정에서는 viterbi training algorithm이 사용된다.

HMM의 인식 문제는 주어진 학습 자료에 대해서는 높은 성능을 보이지만, 학습 자료의 고정된 feature parameter dimension에 대해서만 학습하기 때문에 잡음이나 발화 특성이 다른 음성에 대해서는 인식률이 떨어지는 문제를 보인다. 이 문제를 극복하기 위해, feature parameter dimension을 효율적으로 변화시키며 학습 가능한 심층 신경망(Deep Neural Network, DNN)을 사용하는 방법이 높은 성능을 보이고 있다. DNN은 HMM이 해결할 수 있는 세 가지 문제 중, 인식과 학습을 대체하여 더욱 높은 음성인식률을 보인다. 하지만 강제 정렬에 대해서는 DNN이 HMM을 대체할 수 없다. HMM은 viterbi algorithm을 통해 최적의 상태열을 결정하게 되는데, DNN의 각 상태는 HMM과 달리 특정한 음소를 의미하지 않기 때문에 음성에 대한 인식 단위를 구분할 수 없다. 때문에 일반적인 음향 모델링 방법은 DNN과 HMM을 융합하여 사용하는 hybrid DNN-HMM 모델을 사용한다. HMM을 통해 강제 정렬된 결과를 DNN을 통해 학습하고, DNN의 출력으로 가장 높은 확률을 가지는 HMM의 상태열을 출력하는 방식으로 이루어진다.

Fig. 1은 DNN-HMM 구조를 간략화한 그림이다. 인식을 담당하는 DNN 구조의 출력 확률분포는 기 학습된 HMM 구조로부터 주어진 확률 분포에 의해 결정된다. 즉, 음성이 입력되었을 때, DNN은 가장 높은

확률을 가지는 HMM state의 확률을 출력하게 된다. 출력된 HMM은 모델마다 정의된 음소를 기반으로 음성 인식 결과를 출력한다.

DNN은 목적 함수를 통해 오류를 최소화 시키는 방법으로 최적화를 진행한다. 음성인식에서 주로 사용하는 목적 함수는 최대 우도 추정(Maximum Likelihood Estimation, MLE)이다.

$$\arg_W \max \Sigma_i \log P(o_i|W) = W_{MLE}. \quad (3)$$

Eq. (3) 은 음향 모델에서의 MLE를 나타낸다. o는 관측열 집합의 원소를 나타내며, W는 인식 단위열 집합을 나타낸다. 음향 모델에서 MLE는 어떠한 단어열이 존재할 때 가능한 관측열의 동시 확률을 최대가 되도록 학습을 진행한다. 즉, 음향 모델 학습에서 지향하는 학습 방향을 나타낸다.[3-5]

최대 확률을 가지는 모델을 생성하는 MLE 방식의 학습은 유한한 학습 데이터 내에서의 최대 확률을 구성하는 것이라 다양한 환경의 음성에 대응하기 어렵고, 음성 정보 상호간의 특징을 구별하는 정보를 학습하기 어렵다는 단점이 있다. 상호 정보 최대화(Maximum Mutual Information, MMI)는 학습 데이터의 상호 정보를 최대화하는 방향으로 네트워크를 최적화한다.

$$MMI = \Sigma_u \log \frac{p(o_u|w_u)P(w_u)}{\Sigma_s p(o_u|W)P(W)}. \quad (4)$$

Eq. (4)는 음성인식에서 사용되는 MMI를 나타낸다. 전체 발화 집합 s에 대하여, o는 관측열을 나타내며, w는 학습하고자 하는 발성에 대한 HMM state, W는 w의 열을 나타낸다. 즉, u번째 발화에 대한 발생 확률은 최대화하고, 전체 열에 대한 발생 확률은 최소화하여 상호 정보를 명확하게 나누는 것이 MMI 기반 DNN 학습 방법이다.

음성 인식 문제를 해결하는 데 더 적합한 MMI 방식은 lattice 기반의 MMI를 사용하는 것이다.[6] 일반적으로 음성인식의 출력을 표현할 때 lattice라는 구조를 사용하는데, 이는 어떠한 발성이 입력되었을 때, 이론상 가능한 단어열의 집합을 표현하는 방식이다.

사전에 준비된 lattice가 필요한 lattice-based 방식의 문제점을 해결하고자 제안된 것이 lattice 없는 상호 정보 최대화(Lattice Free Maximum Mutual Information, LF-MMI) 방식이다.[7] 이 방식에서는 단어 기반의 lattice가 아닌 음소 기반의 가중 유한 상태 전이기(weighted Finite State Transducer, wFST)를 사용하여 음향학적 특징과 가까운 범위에서 MMI 식을 사용한다. 이 때, 음소 기반의 wFST를 구성하기 위해 음소 단위의 언어 모델이 사용된다.

음소 단위의 언어모델은 N-gram 방식을 사용한다. 학습 데이터의 정답으로 주어진 단어열들을 발음 사전을 이용하여 발음열로 전환하고, 해당 발음열과의 관계를 N-gram을 이용하여 확률을 계산한다. MMI 식의 사전 확률을 구하기 위해 음소 단위의 언어 모델이 사용된다.

본 논문에서는 MMI 학습에 이용되는 사전 확률의 정확도를 높이기 위해 external knowledge를 음소 기반 언어 모델 학습에 이용한다. 이를 통해 음향 모델은 학습 자료로 주어진 발화 데이터 이외의 knowledge에 대한 영역까지 학습할 수 있다.

본 논문은 다음과 같은 순서로 구성된다. 2장에서는 lattice 기반의 MMI와 LF-MMI에 대한 관련 연구를 소개한다. 3장에서는 본 논문에서 제안하는 external knowledge를 사용한 MMI에 관해 서술한다. 4장에서는 external knowledge를 사용한 MMI에 대한 검증을 위해, 기존 기술과의 음성 인식 성능을 비교 평가한다. 5장에서는 본 논문의 결론을 맺는다.

II. 관련 연구

본 장에서는 external knowledge를 사용한 LF-MMI 기반 음향 모델링 관련 선행 연구를 기술한다. Reference [3]은 MMI 목적식을 HMM 기반의 음성 인식에 적용하였다. Reference [3]에서는 HMM 기반의 음향 모델링을 통해, MLE에 비교하여 MMI의 상대적 인식 오류율이 약 18 % 감소함을 보이며 음향 모델링에서의 MMI의 유용함을 보였다.

Reference [4]에서는 HMM 기반의 음성인식에 대하여, 음소 기반의 TIMIT corpus를 사용하여 실험한 결과를 보였다. Reference [4]에서는, MMI는 MLE에

비해 계산복잡도가 올라가기 때문에 학습 시 연산량이 늘어날 수 있으나, 일정 성능까지 올라가는 데 필요한 매개변수의 개수를 4배 정도 줄일 수 있기 때문에 실시간 음성인식에 있어 반응 속도에 상당한 성능 개선이 있음을 보였다. 따라서, MMI 학습이 MLE 학습에 비해 효율적인 학습을 한다는 것이 증명되었으며, 같은 성능을 보이는 모델에 대해 더 빠른 탐색 속도를 보이는 모델을 만들 수 있음을 증명하였다.

Reference [5]는 음성인식과 같은 구별 문제를 해결할 때, MMI의 조건부 확률로 언어에 대한 확률을 조건부 확률로 반영하여 계산할 때, 성능이 개선됨을 보인다. 이는 음성인식 문제를 해결할 때 MMI를 사용하여 구분 학습을 하는 것이 MLE의 생성 확률을 높이는 목적 함수보다 더욱 적절함을 보인다.

Lattice를 기반으로 한 MMI 방식은 Reference [6]에서 제안되었다. Reference [6]에서는 사전에 만들어진 lattice를 기반으로 DNN 학습을 실시한다. 하나의 음성 프레임에 대해 HMM 생성 확률을 생성하는 기존 DNN 기반 음향 모델과 다르게, 하나의 음성 프레임이 아닌 lattice에 의해 결정된 한 단어의 프레임들의 열 단위로 학습이 이루어진다. 열 정보를 같이 반영하여 학습을 실시했을 때, 그렇지 않은 방법보다 상대적 20 % 성능 향상을 보였다.

Lattice를 기반으로 한 MMI 방식은 사전에 미리 만들어진 lattice를 사용하여 빠르게 음향 모델을 학습할 수 있다는 장점이 있지만, 단어 단위의 lattice를 미리 생성해 놓아야 하기 때문에, 하나의 단어에 해당하는 발성 만큼의 입력이 있어야 결과를 출력할 수 있다. 이는 음향 모델의 출력이 lattice에 있는 단어에 종속되는 문제를 야기한다. Lattice를 미리 생성하지 않는 학습 방법은 계산 복잡도를 낮출 수 있으며, 단어의 제한이 사라지기 때문에 효율적인 학습을 가능하게 한다. 계산 복잡도를 낮추면서 성능을 개선할 수 있는 방법인 lattice-free MMI는 Reference [6]에서 제안되었다. Reference [7]에서는 DNN 모델에서 LF-MMI를 적용하여 학습 최적화 방법을 제안하였다. 특히, Reference [7]에서는 LF-MMI를 적용하여, 강제 정렬을 필요로 하지 않는 DNN 음향 모델 방식을 제안하였다. 즉, 단어별 훈련 자료를 생성하는 것이 아닌, MMI의 음소 모델을 사용하여 음소별 훈련 자료를 생성하여 HMM 훈련 없이 음향 모델을 학습하는 방법을 제안하였다.

III. External knowledge를 이용한 LF-MMI 학습 방법

음향 모델 학습에 있어 MMI가 MLE와 비교하여 가지는 가장 큰 차이점은 목적식에 대한 사전 확률을 가진다는 점이다. 음성 인식에서 사전 확률은 출력에 대한 열 정보이며, 이 확률을 출력하는 모델은 인식 단위의 열로 이루어진 학습 자료를 통해 모델링할 수 있다. 사전 확률 모델은 음향 모델로 하여금 더욱 정확한 확률을 출력할 수 있도록 하고, 이는 학습 속도를 빠르게 할 뿐 아니라 음향 모델의 매개변수를 효율적으로 관리할 수 있도록 한다.[4] 본 논문에서는 MMI학습에 있어 이 사전 확률의 정확도를 더욱 높여 음향 모델의 성능을 개선하는 것에 그 주안점을 둔다.

LF-MMI 기준으로 보았을 때, 음소 모델은 어떠한 음소의 발생 확률에 대한 사전 확률 모델이며, 이와 같은 순서가 있는 데이터를 다룰 때는 N-gram 방식이 가장 널리 쓰이고 있다.

$$P(w_k|w_{k-1} \cdots w_{k-N+1}) . \tag{5}$$

Eq. (5)는 N-gram 방식을 나타낸다. w_k는 어떤 문장에서 k번째 단어를 나타내며, N은 N-gram의 계수를 말한다. 즉, Eq. (5)는 해당 단어와 단어 앞의 N-1개의 단어의 발생 확률을 고려한 단어 확률 생성 모델이라고 볼 수 있다.

N-gram 방식은 graph 형식의 모델이 아닌, 해당하는 단어에 대한 확률값이 사전 형태로 정리되어있는 방식이기 때문에 확률값을 찾을 때 graph 기반의 모델에 비해 탐색 속도가 빠르다는 장점이 있다. 또한, 필요한 모델을 선별적으로 탐색할 수 있다는 장점 또한 존재한다. Reference [7]에서는 LF-MMI 기반 음향 모델 학습에 있어 4-gram 모델을 제안하였다.

사전 확률을 개선하여 음향 모델의 성능을 높이는 방법은, 음성 데이터를 추가하지 않아도 음향에 대한 인식률 향상에 도움이 된다는 장점이 있다. 음소

모델 학습에는 텍스트 데이터를 필요로 하는데, 일반적으로 텍스트 데이터는 음성 데이터에 비해 많은 양을 수집할 수 있기 때문에, 음성 데이터를 늘려서 성능 개선을 하는 것에 비해 효율적인 음향 모델링을 할 수 있다.

기존 LF-MMI에서 음소 모델링은 음향 모델에서 사용하는 정답 텍스트를 기반으로 학습하게 된다. 이 경우, 모델에서 처리 가능한 확률 모델은 음향 모델의 텍스트 정보로 한정된다. 이는 MLE에 비해 빠르게 학습이 진행되는 MMI의 특성 상 과적합을 야기할 수 있으며, 음성 인식률을 하락시킨다.

External knowledge는 자연어 처리 분야에서 기존 모델을 특정 도메인에 적합하게 적응시키거나 다양한 환경에서 테스트가 필요할 때 사용하는 데이터로 학습에 사용하는 데이터 이외의 다른 데이터를 말한다. 음성 인식에서 external knowledge는 언어 모델의 학습 자료로 사용된다. External knowledge는 적응시키고자 하는 분야의 텍스트 데이터를 수집하여 구성된다. 예를 들어, 의료 목적의 음성 인식기를 구축해야 하는 경우, 의료에 관련된 음향 데이터 뿐 아니라 텍스트로 구성된 데이터를 수집하여 언어 모델에 사용함으로써 의학 관련 용어 등을 인식하여 출력할 수 있도록 한다. 음성 인식에서는 구성된 학습 데이터를 N-gram을 통하여 모델링되어 단어의 확률을 생성한다.

본 논문에서 사용되는 External knowledge는 언어 모델에 사용되는 텍스트 데이터를 모두 발음열로 변환하여 LF-MMI 기반 음향 모델 학습에 사용한다. 이를 통해 언어 모델 뿐 아니라 음향 모델에서도 적응시키고자 하는 분야에 대한 발음열의 정보를 같이 학습하여 해당 분야에 대한 인식 성능을 향상시킬 수 있다.

IV. 실 험

본 논문에서 제안하는 방법의 검증을 위하여, 한국어 연속음성인식기를 사용한 음성인식률 평가를 통해 제안한 방법을 평가한다. 평가 대상은 DNN-HMM 기반 음향 모델링에서 최고의 성능을 보이는 LF-MMI 기반 모델과 논문에서 제안하는 external knowledge 기반 모델을 상호 비교하며, 해당 모델을 통해 나온 음성인식 결과의 음절오류율(Character error rate, CER)을 서로 비교한다.

Table 1. The training data for acoustic model.

Training data	Num. of utterance	Hours (Hr.)
ETRI Korean reading DB	100,000	277.78
SiTEC Korean reading DB 01	20,806	57.79
Korean mobile assistant DB	92,874	100.00
Test data	Num. of utterance	Hours (Hr.)
Korean mobile assistant DB for test	876	1.35

Table 1은 본 연구에 사용된 훈련 데이터 및 테스트 데이터를 나타낸다. 실험에 사용된 데이터는 한국어 데이터로 구성되어 있으며, 조용한 환경에서 한국어로 대화한 320시간의 16,000 Hz sapling rate를 가진 말뭉치(corpus)로 구성되어 있다. 여기에 카페, 버스, 기차 등 일상 생활의 배경 잡음을 SNR(Signal-to-Noise Ratio) 2에서 6 사이의 값을 랜덤하게 적용한 320시간 분량의 말뭉치를 더해 총 640시간의 말뭉치를 사용한다. 여기에 휴대폰에서 녹음된 음성인식 비서 도메인의 말뭉치를 추가하여 총 740시간의 학습 데이터를 음향 모델에 사용한다. 테스트 자료는 잡음이 섞이지 않는, 마이크로 녹음된 음성인식 비서 도메인을 사용하였다. 언어 모델 의 경우, 웹 크롤링을 통해 자체 수집한 3억 7천만 문장분량의 한국어 말뭉치를 사용하였다. 음소 모델로 사용되는 external knowledge는 해당 언어 모델을 음소로 변환한 결과를 사용하여 사전 확률을 계산하였다.

음향 모델의 전처리와 학습은 Kaldi toolkit을 사용하여 진행하였다.[8] 이는 현재 LF-MMI를 지원하는 음성인식 학습 toolkit 중 가장 높은 성능을 보이며, 대용량 처리가 가능하고, 완성된 모델을 조합하는 wFST 기반의 decoding network를 지원한다.

음향 모델의 입력과 출력은 각각 다음과 같이 정의한다. 음향 모델의 입력은, 오디오 신호로부터 추출한 40차의 필터뱅크의 값으로 나타낸다. 이는 오디오 신호에서 음성의 특징을 추출하여 인식 성능을 높이기 위함이며, 음성을 25 ms 구간의 프레임으로 분리한 뒤 일정 주파수 영역의 값에 필터를 적용한

Table 2. Acoustic model hyperparameter.

Model	Num. of hidden layer	Num. of node for layer
TDNN	11	1,280

값이다. 필터뱅크 방식은 인간의 청각 기관을 모방한 방법으로, 음성 인식에 있어 가장 널리 쓰이는 특징 추출 방법이다. 본 논문에서는 해당 도메인에서 가장 높은 성능을 보이는 40차 필터뱅크를 사용하였다.[9-10] 음향 모델의 출력은 학습된 HMM 모델로부터 얻어진 forced-alignment된 HMM 모델의 결과를 사용하여 학습을 진행한다. 따라서, DNN-HMM 기반의 음향 모델에서의 출력은 해당 확률을 갖는 HMM state의 확률이 된다.

DNN을 사용하는 음향 모델 중, 현재 가장 높은 성능을 보이는 모델은 TDNN(Time Delayed Neural Network)이다. 음성인식 문제에서 DNN이 하는 역할은 입력으로 주어진 하나의 프레임에 대하여 가장 높은 확률을 가지는 HMM 모델의 발생 확률을 구분해 주는 것이다. 또한 음성은 시간 종속적인 특성을 지니고 있어, DNN의 발생 확률을 구분함에 있어 시간 종속적인 특성을 지닐 수 있도록 DNN에 앞, 뒤 프레임을 같이 넣어줌으로써 시간 종속적인 특성을 반영한다. 이러한 시간 종속적인 도메인에 대해 효율적인 학습을 가능하게 하는 모델이 TDNN이다. TDNN은 하나의 node가 앞 layer의 모든 node에 연결되어 있는 기존 DNN과 달리 하나의 node가 앞 layer의 일정 구간의 node에만 연결되어 있어, 해당 node는 특정 구간에 대해서만 학습을 할 수 있게 해 준다. 본 논문에서는 TDNN을 활용하여 실험을 진행하였다.

Table 2는 음향 모델에 사용된 TDNN을 구성하는 hyperparameter를 나타낸다. 11개의 은닉층으로 이루어져 있으며, layer당 node의 개수는 1,280개이다. 입력은 40차 fbank의 앞, 뒤 2개씩의 context를 합친 총 200차의 node를 사용하였으며, 출력은 HMM으로부터 출력된 6,528개가 사용되었다.

External knowledge는 언어 모델 학습에 사용된 3억 7천만 문장의 text corpus를 그대로 사용하였으며, 이를 국립국어원에서 공표한 표준어 규정 중 제 2부인 표준 발음법을 참조하여 음소로 변환하였다.[11]

Table 3은 MLE 기반의 DNN과 TDNN, LF-MMI 기반의 TDNN과 external knowledge를 사용한 음향 모델의 성능 비교를 나타낸다. DNN 기반의 음성인식기의 음절오류율은 3.58 %를 나타내었고, 이 모델을 TDNN으로 변경한 결과 1.97 %로 1.62 %의 절대적 성능 향상을 보였다. 또한 목적식을 LF-MMI로 변경한 결과, 1.56 %로 MLE 기반의 TDNN과 비교했을 때 0.41 %의 절대적 성능 향상을 보였다. 또한, external knowledge를 사용한 LF-MMI를 적용했을 때 1.34 %로 MLE를 사용한 TDNN과 비교하여 0.63 %의 절대적 성능 향상을 보였다.

Table 3. Experimental result for acoustic models and their objective functions.

Acoustic model	CER (%)
DNN+MLE	3.58
TDNN+MLE	1.97
TDNN+LF-MMI	1.56
TDNN+external knowledged LF-MMI	1.34

V. 결 론

본 논문은 external knowledge를 사용하여 DNN의 목적식을 LF-MMI로 사용한 음향 모델링 방법을 제안한다. LF-MMI는 DNN 학습에서 널리 쓰이는 방식인 MLE 방식과 비교하여 구별 학습의 영역에서 높은 성능을 보인다.

본 논문에서는 LF-MMI의 사전 확률을 담당하는 음소 모델에 external knowlege를 사용함으로써 기존 방식의 단점인 음향 모델의 성능을 높이는 방법을 제안한다. HMM에서 얻어낸 forced-alignment값을 통해 DNN학습을 수행했을 시, 기존 LF-MMI와 비교하여 14 %의 상대적 성능 개선을 보였다.

감사의 글

본 연구는 산업통상자원부의 산업기술혁신사업으로부터 지원을 받아 수행된 연구임(No. 10063424, ‘실내용 음성대화 로봇을 위한 원거리 음성인식 기술 및 멀티 태스크 대화처리 기술 개발’).

References

1. B. Juang and L. Rabiner, “Hidden Markov models for speech recognition,” Technometrics, **33**, 251-272 (1991).
2. S. Suwon, J. Rho, S. Kim, J. Lee, and H. Ko, “Text independent speaker verficiation using dominant state information of HMM-UBM,” J. Acoust. Soc. Kr. **34**, 171-176 (2015).
3. L. Bahl, P. Brown, P. de Souza, and R. Mercer, “Maximum mutual information estimation of hidden Markov model parameters for speech recognition,” Proc. ICASSP, **11**, 49-52, (1986).
4. S. Kapadia, V. Valtchev, and S. Young, “MMI training for continuous phoneme recognition on the TIMIT database,” Proc. ICASSP, **2**, 491-494 (1993).
5. D. Yu and L. Deng, *Automatic Speech Recognition* (Springer London limited, London, 2016), pp. 193-215.
6. B. Kingsbury, “Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling,” Proc. ICASSP, **2**, 3761-3764 (2009).
7. D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, “Puerly sequence-trained neural networks for ASR based on lattice-free MMI,” Proc. Intetspeech, 2751-2755 (2016).
8. D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, “The Kaldi speech recognition toolkit,” Proc. ASRU. (2011).
9. H. Hadian, H. Sameti, D. Povey, and S. Khudanpur, “End-to-end speech recognition using lattice-free MMI,” Proc. Interspeech, 2345-2349 (2013).
10. H. Lim, M. Kim, and H. Kim, “Sound event classification using deep neural network based transfer learning,” (in Korean), J. Acoust. Soc. Kr. **35**, 143-148 (2016).
11. K. Lee, J. Jeon, and M. Jung, “Automatic generation of pronunciation variants for Korean continuous speech recognition” (in Korean), J. Acoust. Soc. Kr. **20**, 35-43 (2001).

저자 약력

▸ 박 호 성 (Hosung Park)

2016년 2월 : 한동대학교 전산전자공학부 학사
2018년 2월 : 서강대학교 컴퓨터공학과 석사
2018년 3월 ~ 현재 : 서강대학교 컴퓨터공학과 박사과정

▸ 강 요 셉 (Yoseb Kang)

2017년 2월 : 서강대학교 수학과/경제학과 학사
2019년 2월 : 서강대학교 컴퓨터공학과 석사
2019년 3월 ~ 현재 : NCSOFT Speech lab 음성인터페이스팀 연구원

▸ 임 민 규 (Minkyu Lim)

2008년 2월 : 서강대학교 기계공학과/컴퓨터공학과 학사
2010년 2월 : 서강대학교 컴퓨터공학과 석사
2019년 8월 : 서강대학교 컴퓨터공학과 박사
2019년 9월 ~ 현재 : SK텔레콤 음성인식기술 cell manager

▸ 이 동 현 (Donghyun Lee)

2013년 2월 : 서강대학교 컴퓨터공학과 학사
2013년 3월 ~ 현재 : 서강대학교 컴퓨터공학과 석박사통합 과정

▸ 오 준 석 (Junseok Oh)

2017년 8월 : 서강대학교 컴퓨터공학과 학사
2019년 8월 : 서강대학교 컴퓨터공학과 석사
2019년 9월 ~ 현재 : 서강대학교 청각지능 연구실 연구원

▸ 김 지 환 (Ji-Hwan Kim)

1996년 2월 : KAIST 전산학과 학사
1998년 2월 : KAIST 전산학과 석사
2001년 11월 : Cambridge University Engineering Department 박사
2007년 8월 : LG전자 책임연구원
2007년 9월 ~ 현재 : 서강대학교 컴퓨터공학과 교수