한국음향학회지 제39권 제5호 pp. 483~489 (2020)
The Journal of the Acoustical Society of Korea Vol.39, No.5 (2020)
https://doi.org/10.7776/ASK.2020.39.5.483

pISSN : 1225-4428
eISSN : 2287-3775

저자원 환경의 음성인식을 위한 자기 주의를 활용한 음향 모델 학습

Acoustic model training using self-attention for low-resource speech recognition

박호성,[1] 김지환[1†]

(Hosung Park[1] and Ji-Hwan Kim[1†])

[1]서강대학교 컴퓨터공학과

(Received August 7, 2020; accepted September 4, 2020)

초 록: 본 논문에서는 저자원 환경의 음성인식에서 음향 모델의 성능을 높이기 위한 음향 모델 학습 방법을 제안한다. 저자원 환경이란, 음향 모델에서 100시간 미만의 학습 자료를 사용한 환경을 말한다. 저자원 환경의 음성인식에서는 음향 모델이 유사한 발음들을 잘 구분하지 못하는 문제가 발생한다. 예를 들면, 파열음 /d/와 /t/, 파열음 /g/와 /k/, 파찰음 /z/와 /ch/ 등의 발음은 저자원 환경에서 잘 구분하지 못한다. 자기 주의 메커니즘은 깊은 신경망 모델로부터 출력된 벡터에 대해 가중치를 부여하며, 이를 통해 저자원 환경에서 발생할 수 있는 유사한 발음 오류 문제를 해결한다. 음향 모델에서 좋은 성능을 보이는 Time Delay Neural Network(TDNN)과 Output gate Projected Gated Recurrent Unit (OPGRU)의 혼합 모델에 자기 주의 기반 학습 방법을 적용했을 때, 51.6 h 분량의 학습 자료를 사용한 한국어 음향 모델에 대하여 단어 오류율 기준 5.98 %의 성능을 보여 기존 기술 대비 0.74 %의 절대적 성능 개선을 보였다.
핵심용어: 음성 인식, 음향 모델, 자기 주의 메커니즘, 저자원 환경

ABSTRACT: This paper proposes acoustic model training using self-attention for low-resource speech recognition. In low-resource speech recognition, it is difficult for acoustic model to distinguish certain phones. For example, plosive /d/ and /t/, plosive /g/ and /k/ and affricate /z/ and /ch/. In acoustic model training, the self-attention generates attention weights from the deep neural network model. In this study, these weights handle the similar pronunciation error for low-resource speech recognition. When the proposed method was applied to Time Delay Neural Network-Output gate Projected Gated Recurrent Unit (TNDD-OPGRU)-based acoustic model, the proposed model showed a 5.98 % word error rate. It shows absolute improvement of 0.74 % compared with TDNN-OPGRU model.
Keywords: Speech recognition, Acoustic model, Self-attention mechanism, Low-resource environment

PACS numbers: 43.72.Bs, 43.72.Ne

I. 서 론

인간과 컴퓨터의 의사 소통 방법 중, 음성인식은 가장 직관적인 방법 중 하나이다. 음성인식이란, 사람이 발성하는 음성을 컴퓨터가 마이크 등으로 수신하여 문자로 해석하는 것을 말한다. 음성인식 기술은 자동차 네비게이션, 가정용 스마트 스피커 등 인간과 컴퓨터의 소통이 필요한 다양한 분야에서 응용되고 있다.

음성인식의 구현을 위해서는 음향 모델과 언어 모델, 발음 사전과 디코딩 네트워크를 필요로 한다. 음향 모델은 학습 자료를 토대로 음성을 입력 받아 가

†**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)

장 높은 확률을 가지는 음소의 열을 출력하는 모델을 말하며, 언어 모델은 단어의 열이 주어졌을 때, 학습 자료를 토대로 다음에 나올 수 있는 단어의 확률을 출력하는 모델이다. 발음 사전은 음소와 단어의 관계성을 정의해놓은 사전을 말하며, 디코딩 네트워크는 음향 모델과 언어 모델, 발음 사전을 묶어 하나의 음성 인식 과정을 수행할 수 있는 네트워크를 말한다. 이 중, 음향 모델은 음향 정보를 입력 받아야 하는 음성인식 과정에서 가장 중요한 역할을 담당하고 있으며, deep neural network과 hidden Markov model을 이용한 DNN-HMM 음향 모델링 방법이 가장 높은 성능을 보이고 있다.[1]

DNN-HMM 기반 음향 모델은 입력으로 주어지는 특징 벡터에 대해 발음 사전에 정의된 음소 중, 가장 높은 확률을 가지는 음소를 출력하는 분류 과정을 수행한다. 음향 모델의 분류 성능을 높이기 위한 방법은 두 가지이다. 첫 번째는 입력으로 주어진 단일 특징 벡터에 대한 분류 성능을 높이는 과정이다. State-level minimum Bayes risk(sMBR), Maximum Mutual Information(MMI) 등의 강력한 분류 성능을 가진 DNN 목적 함수가 사용되며, 이를 이용하여 구별 학습을 수행한다.[2] 두 번째는 음성 인식은 시간 종속적인 벡터의 열을 추정하는 문제이므로, 단일 특징 벡터가 아닌 특징 벡터의 열을 입력으로 받아 가장 높은 확률을 가지는 음소의 열을 출력하는 방법이다. 이 방법에서는 Recurrent Neural Network(RNN), Long-Short Term Memory(LSTM),[3] Gated Recurrent Unit (GRU),[4] Time Delay Neural Network(TDNN)[5] 등의 DNN 모델을 사용하여 열을 학습한다.

그러나, 음향 모델 기준 100 h 미만의 음성 데이터를 사용한 저자원 환경의 음향 모델 학습에서는 DNN 모델이 충분히 학습되지 못해 음향 모델의 분류 성능에 악영향을 끼치는 과적합 및 과소적합 문제가 발생할 수 있다.[6] 분류 성능에 악영향을 끼치는 대표적인 문제는 유사한 발음들에 대한 분류 오류이다. 유사한 발음이란, 동일 조음위치를 가진 자음, 모음에 대한 발음 및 발화자의 발성 습관 등에 따라 모호함이 발생할 수 있는 발음들을 말한다. 예를 들어, '금연'과 '크면'의 첫 발성인 /g/와 /k/, '두자'와 '투자'의 첫 발성인 /d/와 /t/, '회'와'훼'(/ø/와 /we/), '개'와 '게'(/ɛ/와 /e/) 등의 발음을 유사한 발음이라 할 수 있다. 음향 모델에 대해 충분한 양의 학습데이터가 주어진 경우, 유사한 발음 문제는 하나의 발성 단어에 대한 다양한 형태의 데이터를 학습 할 수 있기 때문에 위에 언급한 음향 모델 학습의 두 번째 방법인 특징 벡터의 열을 입력으로 받는 방법으로 극복 가능하다. 하지만, 저자원 환경에서는 충분한 양의 데이터가 주어지지 않아 유사한 발음들에 대한 문제가 발생하고 있다.

본 논문은 저자원 환경에서의 유사한 발음들을 잘 구분하기 위하여 열을 다룰 수 있는 GRU 모델에 자기 주의 메커니즘을 적용하는 유사한 발음에 대해 높은 성능을 보이는 구분 벡터 학습을 제안한다. 구분 벡터 학습이란 DNN-HMM 기반 음향 모델의 은닉층에서 출력되는 다차원 벡터를 입력 받아 가중치를 부여한 뒤 동일 차원의 벡터로 출력하여 분류 성능을 향상시키는 학습 방법이다. 본 논문에서는 자기 주의 메커니즘을 사용하여 가중치를 부여하였으며,[7] 51.6 h 분량의 저자원 한국어 학습 자료에 대해 기존 음향 모델과 비교하여 높은 유사한 발음 분류 성능을 보였다.

본 논문은 다음과 같은 순서로 구성된다. 2장에서는 DNN-HMM 기반 음향 모델에서 높은 성능을 보이는 TDNN-Output gate Projected Gated Recurrent Unit (TNDD-OPGRU) 기반의 음향 모델 및 self-attention mechanism에 대해 다룬다. 3장에서는 본 논문에서 제안하는 self-attention을 활용한 discriminative vector 학습에 대해 다룬다. 4장에서는 기존 기술과의 비교 평가를 통해 한국어 저자원 데이터에 대한 제안한 모델을 검증한다. 5장에서는 본 논문의 결론을 서술한다.

II. 관련 연구

DNN-HMM 기반의 음향 모델 방법 중, 가장 높은 성능을 보이는 모델은 TDNN[5]과 RNN의 혼합 모델이다. 특히, RNN의 경우는 LSTM 및 GRU를 사용하는데,[4] 이는 출력에 비해 입력의 길이가 긴 음향 모델의 특징 때문이다. 음성 인식의 특징 벡터는 10 ms 간격으로 20 ms ~ 30 ms 길이의 발성을 대상으로 하기 때문에, 5 s 동안 20개의 음소를 발성하는 경우를 예로

들었을 때, 이 발성에 대해서 20개의 음소에 대해 500개의 벡터가 생성된다. 이 때문에 발생하는 RNN의 기울기 소실문제를 해결하기 위해 LSTM이나 GRU가 사용된다.[4]

이 중, GRU의 학습 효율을 높인 output gate projected GRU(OPGRU)가 RNN은 사용하는 음향 모델 중 가장 높은 성능을 보이고 있다.[8] GRU는 선행 벡터 정보의 적용 여부를 판단하는 reset gate와, 현재 벡터 정보를 저장하기 위한 update gate로 이루어져 있는데, 이를 음향 모델 학습에 응용할 시 reset gate의 값이 너무 일찍 수렴하는 문제가 발생한다.[8] 이를 해결하기 위하여, Reference [8]에서는 기존 GRU의 입력 부분에 위치했던 reset gate의 위치를 GRU의 출력 부분으로 옮김으로서 계산 복잡도를 유지한 채 학습 효율을 높이는 방법인 OPGRU를 제안하였다.

주의 메커니즘은 임의 길이의 입력 벡터의 열이 입력으로 주어졌을 때, 벡터들의 정보를 효과적으로 압축하여 전달하기 위해 제안된 방법이다.[9] 자기 주의 메커니즘은 이를 응용하여 병렬 학습이 가능하도록 제안된 모델이며,[10] 이후 정렬된 입력에 대하여 주의 메커니즘을 적용하는 local attention이 제안되었다.[11] 음성인식에서는 Reference [11]의 연구를 확장한 time-restricted self-attention이 제안되었으며,[12] 이 방법은 고정된 길이에 대한 벡터 가중치를 연산하는 방법으로, 음성인식의 입력 벡터의 차원 수가 일정하다는 점에 착안한다.

III. Self-attention mechanism을 이용한 discriminative vector 학습

본 논문에서는 discriminative vector 학습을 위해 Reference [11]에서 제안하는 self-attention layer를 사용하였으며, 구현에서 사용된 활성함수 및 학습 방법은 Reference [12]의 time-restricted self-attentive layer 방식을 사용하여 Fig. 1와 같이 구현하였다.

Self-attention mechanism은 하나의 입력에 대해 3가지 요소들을 통해 입력을 처리한다. 각각은 query, key, value로 정의된다. Self-attention에서는 하나의 입력에 대하여 고정된 크기의 서로 다른 query와 key, value 행렬을 곱한 값을 통해 query, key, value의 값을

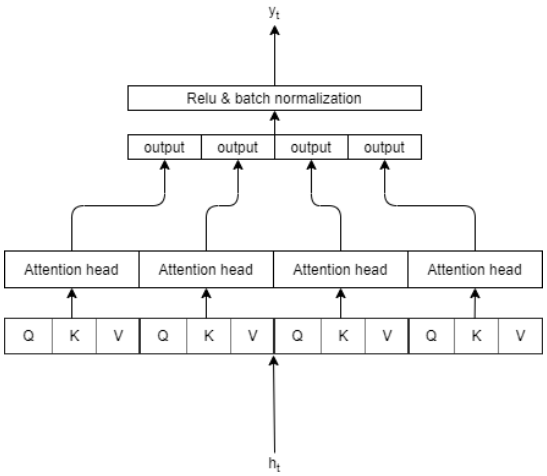

Fig. 1. Multi-head self-attention mechanism.

생성한다. Self-attention은 query와 key 사이의 cosine 유사도를 계산하는데, 이것을 주의 가중치라 한다. 이 주의 가중치와 value와의 element-wise multiplication 및 softmax 함수를 통한 정규화를 통해 최종 출력이 나오게 된다. 이 일련의 과정이 수행되는 고정된 구간을 하나의 head라고 하며, 본 논문에서는 Reference [11]에서 제안하는 여러 개의 head를 사용하는 방법인 multi-head attention을 사용한다.

본 논문에서는 TDNN-OPGRU 모델의 출력 벡터를 대상으로 self-attention layer를 적용하여 구분 벡터 학습을 수행한다. TDNN-OPGRU 모델의 구현은 Reference [8]을 참고하여 진행하였으며, 실험하고자 하는 데이터에 맞게 hyperparameter를 재구성하여 구현하였다.

TDNN은 입력된 모든 벡터들에 대해 모든 노드가 연결되어 있는 일반적인 DNN과 다르게 하나의 노드는 고정된 구간만큼의 벡터들을 입력받는다. 해당 작업을 계층을 쌓아 반복하면, 최상위 레이어에서는 전체 구간에 대한 압축된 정보를 받을 수 있다.[5] 음성 정보의 특징은 인접한 벡터들은 긴밀한 관계를 가지나, 입력 시간의 차이가 클수록 관계성이 약화되는 특징을 가지고 있다. 예를 들면, 10개의 음소를 발성한다고 했을 때, 첫 번째 음소와 두 번째 음소는 음향학적으로 긴밀한 관계를 가지나, 첫 번째 음소와 열 번째 음소는 관계성이 거의 없다. 이 때문에 음향 모델에서는 TDNN과 같은 DNN 모델이 가장 효율적으로 사용된다. GRU 모델은 음성의 시간 종속적인

정보를 모델링하기 위해 사용하며, reset gate[8]의 위치를 이동시켜 효율적인 학습을 가능하게 한 OPGRU 모델을 사용한다.

Fig. 2은 OPGRU의 구조를 설명한다. 이전 시점의 정보는 점선으로 표현되었으며, 현재 시점 **t**의 정보는 실선으로 표현된다. 현재 시점의 벡터 x_t는 이전 시점의 출력 정보를 나타내는 h_{t-1} 와의 곱셈 연산을 수행한 뒤, tanh로 활성화하여 $\underline{h}$를 출력한다. 출력된 $\underline{\boldsymbol{h}}$는 update gate $\boldsymbol{z}$를 통과하여 이전 시점의 출력 정보인 h_{t-1}와의 합연산을 수행한 결과를 output gate에 적용하여 element-wise 곱셈 연산 ⊗을 수행하고 특정 차수의 벡터로 압축하기 위한 투영과정을 수행한다.

음향 모델의 출력 벡터는 self-attention layer를 통과하여 가중치가 부여된 결과값을 softmax 함수를 통해 활성화하여 가장 높은 확률을 가진 음소의 열을 출력한다.

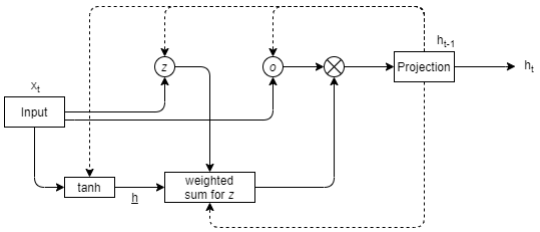

Fig. 2. Output gate GRU architecture.

Fig. 3은 TDNN-OPGRU구조에 self-attention layer를 추가한 음향 모델의 구조를 나타낸다. Self-attention layer는 TDNN과 OPGRU의 혼합 모델의 출력 벡터를 입력받아 가중치를 부여하며, softmax 함수를 통해 가장 높은 확률을 가지는 HMM state의 확률 분포를 출력한다.

IV. 실 험

제안하는 방법에 대한 검증을 위하여, 저자원 음성 말뭉치인 Zeroth 한국어 데이터를 사용하였다.[13] Zeroth 한국어 데이터의 구성은 51.6 h의 음성 및 전사 학습 자료와 1.2 h의 평가용 음성 및 전사 자료로 구성되어 있다.

검증을 위한 실험에 사용되는 발음 사전과 언어 모델은 Zeroth 한국어 데이터에 포함된 발음 사전과 언어 모델을 사용한다. 포함된 언어 모델은 SRILM toolkit[14]을 이용해 학습된 ARPA format 형태로 제공되며, 3-gram 및 4-gram으로 학습 된 두 개의 모델을 제공한다. 3-gram은 음향 모델과 결합한 디코딩 네트워크를 구성하는 데 사용되며, 4-gram으로 학습된 모델은 음성 인식 결과의 re-scoring에 활용된다.[15]

음성의 특징 추출은 40차의 Mel-Frequency Cepstral

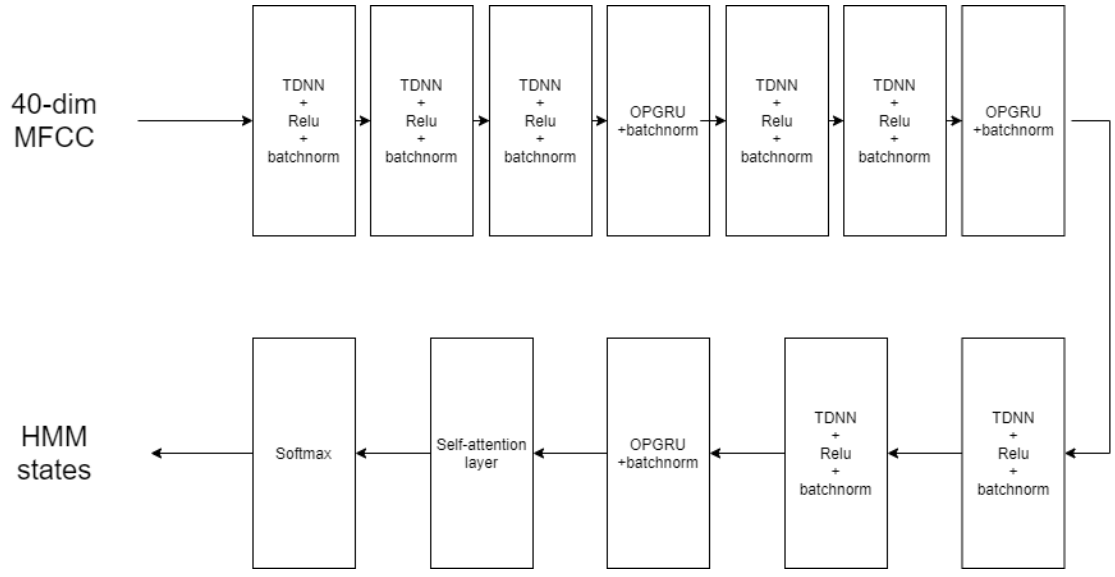

Fig. 3. TDNN-OPGRU-based acoustic model using self-attention layer.

Table 1. Word error rates for the number of self-attention head.

Model		Number of head						Baseline (WER,%)
		40 (WER, %)	60 (WER, %)	120 (WER, %)	150 (WER, %)	180 (WER, %)	200 (WER, %)	
TDNN+attention	4-gram rescoring	10.26	10.17	9.87	9.83	9.78	9.86	10.55
	3-gram	16.07	15.94	16.19	15.61	15.71	15.61	17.65
TDNN+OPGRU+ attention	4-gram rescoring		9.45	9.13	8.25	8.38		9.45
	3-gram		15.22	14.15	12.31	14.09		15.22

Table 2. Word error rate for key/value dimension pair.

Model		Key/value dimension pair				
		20/80 (WER, %)	40/60 (WER, %)	50/50 (WER, %)	60/40 (WER, %)	80/20 (WER, %)
TDNN+OPGRU +attention (Num. head=150)	4-gram rescoring	8.24	8.25	8.49	8.25	8.13
	3-gram	12.24	12.31	12.45	12.1	12.22

Coefficient(MFCC)를 사용한다. 또한, 16,000의 샘플링 레이트를 가지는 음성 신호에 대하여 Nyquist 정리에 의해 복원된 8,000의 샘플링 레이트의 대역 중, 20 Hz ~ 7,600 Hz 범위에 대한 값만 취하는 high-resolution 기술을 적용한 40차의 MFCC 벡터를 음성 특징 벡터로 사용한다.[12]

본 논문의 음향 모델 학습에 사용되는 데이터 증강 방식은 두 가지이다. 첫 번째는 speed perturbation으로 불리는 방식이며, 입력된 음성을 서로 다른 속도로 재생하여 데이터를 증강시키는 방법이다. 본 실험에서는 재생 속도의 0.8배, 0.9배, 1.1배, 1.2배의 재생 속도를 가진 데이터를 추가하여 학습 자료로 사용하였다. 두 번째는 SpecAugment를 사용한 방법이다.[16] 구현은 Reference [16]을 참고하여 진행했으며, speed perturbation의 결과에 적용해 학습 자료로 사용하였다.

본 실험에서의 평가 척도는 Word Error Rate(WER)을 사용하였다. WER은 하나의 문장에서 정답으로 주어진 단어 단위의 전사 자료와 비교하여 음성인식 결과에 추가된 단어, 제거된 단어, 대체된 단어의 개수를 더해 정답 단어의 개수로 나눈 값을 백분율로 나타낸다.[17]

Table 1은 제안한 자기 주의 메커니즘을 적용할 시, 가장 높은 성능을 보이는 head의 개수를 찾는 작업이다. TDNN과 OPGRU의 혼합 모델의 경우, 150개의 head를 사용하는 것이 가장 낮은 단어 오류율을 보임을 확인할 수 있으며, TDNN 단일 모델의 경우 180개의 head를 사용하는 것이 가장 높은 성능을 보임을 알 수 있다.

Table 3. Experiment results compared with baseline acoustic models.

Model	WER (%)
TDNN	10.55
TDNN+OPGRU	9.45
TDNN+OPGRU+SpecAugment (baseline)	6.72
TDNN+self-attention	9.78
TDNN+OPGRU+self-attention	8.13
TDNN+OPGRU+SpecAugment+self-attention (proposed)	5.98

Table 2는 제안한 자기 주의 메커니즘에서 가장 높은 성능을 보이는 사용되는 key/value 쌍을 찾기 위한 실험 결과이다. TDNN과 OPGRU의 혼합 모델의 경우, key/value의 쌍이 80/20의 dimension을 가질 때 가장 높은 성능을 보임을 확인할 수 있다. 동일 연산량에서 가장 낮은 단어 오류율을 보임을 확인하기 위해, 모든 dimension의 합이 일정한 환경에서 실험을 진행하였으며, 실험 결과 key dimension이 높을수록 높은 성능을 보임을 확인할 수 있었다.

Table 3은 제안한 음향 모델에 대한 성능 비교를 나타낸다. 기존 연구에서 사용된 TDNN 기반의 음향 모

Table 4. Example errors compared with baseline acoustic model.

Model	Example sentences
Answer	ex1: ...사랑을 제대로 못 받고 크면... ex2: 여호와의 자비가...
TDNN+OPGRU +Specaugment	ex1: ...사랑을 제대로 못 받고 **금연**... (/g/, 연구개음) ex2: 여호와의 **차비**가 ... (/ch/, 경구개음)
TDNN+OPGRU +Specaugment +self-attention	ex1: ...사랑을 제대로 못 받고 **크면**... (/k/, 연구개음) ex2: 여호와의 **자비**가 ... (/z/, 경구개음)

델에 제안한 방법을 추가한 결과, 기존 대비 0.77 %의 단어 오류율 감소를 보였다. 또한 TDNN과 OPGRU의 혼합 모델에 제안한 방법을 적용했을 때, 1.32 %의 개선을 보였다. 결과적으로, 기존의 성능인 9.45 %의 단어 오류율과 비교하여, 제안한 모델은 5.98 %의 단어 오류율을 보이며 3.57 %의 절대적 성능 개선을 보였다.

Table 4는 제안한 방법을 적용했을 때의 문제 해결 결과 예시를 나타낸다. ex1, ex2와 같이, 기존 모델에서 해결하지 못했던 동일한 연구개음에서 발생하는 오류가 제안된 방법에서는 해결됨을 볼 수 있다. 기존 모델에서 발생한 단어 오류는 전체 평가 데이터의 단어 수인 9,253개의 단어 중 622개이며, 이 중 본 논문에서는 622개 중 69개의 오류를 해결하였다. 69개의 오류 중 38개의 오류가 논문에서 제시한 문제인 유사한 발음의 오류이다. 이는 DNN 벡터에 대한 가중치를 부여한 구분 벡터 학습이 성공적으로 문제를 해결함을 보인다.

V. 결 론

본 논문은 유사한 발음에 대한 한국어 음향 모델의 성능을 개선하기 위해 자기 주의 계층을 적용하여 성능을 향상시킨 음향 모델을 제안하였다. TDNN-OPGRU 기반 음향 모델의 최상위 계층에 자기 주의를 적용하여 실험했을 때, 5.98 %의 단어 오류율을 보였으며, 특히 동일 조음 위치에서 발생하는 자음들에 대한 인식 성능이 개선되어 기존 기술 대비 0.74 %의 성능 개선을 보였다.

감사의 글

이 논문은 2017년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임(No.2017-0-01772, 비디오 튜링 테스트를 통과할 수준의 비디오 스토리 이해 기반의 질의응답 기술 개발).

References

1. C. Weng and D. Yu, “A comparison of lattice-free discriminative training criteria for purely sequence-trained neural network acoustic models,” Proc. ICASSP. 6430-6434 (2019).
2. W. Michel, R. Schluter, and H. Ney, “Comparison of lattice-free and lattice-based sequence discriminative training criteria for LVCSR,” arXiv:1907.01409 (2019).
3. J. Jorge, A. Gimenez, J. Iranzo-Sanchez, J. Civera, A. Sanchis, and A. Juan, “Real-time one-pass decoder for speech recognition using LSTM language models,” Proc. Interspeech, 3820-3824 (2019).
4. J. Y. Chung, C. Gulcehre, K. H. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” arXiv:1412.3555 (2014).
5. V. Peddinti, D. Povey, and S. Khudanpur, “A time delay neural network architecture for efficient modeling of long temporal contexts,” Proc. Interspeech, 2-6 (2015).
6. B. Christian and T. Griffiths, *Algorithms to Llive by: The Computer Science of Human Decisions Chapter 7: Overfitting* (William Collins, Hampshire, 2017), pp. 149-168.
7. D. Povey, H. Hadian, P. Ghahremani, K. Li, and S. Khudanpur, “A time-restricted self-attention layer for ASR,” Proc. ICASSP. 5874-5878 (2018).
8. G. Cheng, D. Povey, L. Huang, J. Xu, S. Khudanpur, and Y. Yan, “Output-gate projected gated recurrent unit for speech recognition,” Proc. Interspeech, 1793-1797 (2018).
9. D. Bahdanau, K. H. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” Proc. ICLR. 1-15 (2015).
10. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” Proc. NIPS. 5999-6009 (2017).
11. M.-T. Luong, H. Pham, and C. D. Manning, “Effective approaches to attention-based neural machine translation,” Proc. EMNLP. 1412-1421 (2015).

12. D. Povey, H. Hadian, P. Ghahremani, K. Li, and S. Khudanpur, “A time-restricted self-attention layer for ASR,” Proc. ICASSP. 5874-5878 (2018).
13. *Zeroth Korean*, http://openslr.org/40/, (Last viewed June 4, 2020).
14. A. Stolcke, “SRILM-an extensible language modeling toolkit,” Proc. ICSLP. 901-904 (2002).
15. H. Xu, T. Chen, D. Gao, Y. Wang, K. Li, N. Goel, Y. Carmiel, D. Povey, and S. Khudanpur, “a pruned RNNLM lattice-rescoring algorithm for automatic speech recognition,” Proc. ICASSP. 5929-5933 (2018).
16. D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, “SpecAugment: a simple data augmentation method for automatic speech recognition,” Proc. Interspeech, 2613-2617 (2019).
17. Y.-Y. Wang, A. Acero, and C. Chelba, “Is word error rate a good indicator for spoken language understanding accuracy,” Proc. ASRU. 577-582 (2003).

저자 약력

▸ 박 호 성 (Hosung Park)

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

▸ 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