한국음향학회지 제41권 제3호 pp. 335~341 (2022)
The Journal of the Acoustical Society of Korea Vol.41, No.3 (2022)
https://doi.org/10.7776/ASK.2022.41.3.335

pISSN : 1225-4428
eISSN : 2287-3775

Sequence dicriminative training 기법을 사용한 트랜스포머 기반 음향 모델 성능 향상

Improving transformer-based acoustic model performance using sequence discriminative training

이채원,[1] 장준혁[2†]

(Chae-Won Lee[1] and Joon-Hyuk Chang[2†])

[1]한양대학교 융합전자공학과

(Received March 21, 2022; revised May 9, 2022; accepted May 9, 2022)

초 록: 본 논문에서는 기존 자연어 처리 분야에서 뛰어난 성능을 보이는 트랜스포머를 하이브리드 음성인식에서의 음향모델로 사용하였다. 트랜스포머 음향모델은 attention 구조를 사용하여 시계열 데이터를 처리하며 연산량이 낮으면서 높은 성능을 보인다. 본 논문은 이러한 트랜스포머 AM에 기존 DNN-HMM 모델에서 사용하는 가중 유한 상태 전이기(weighted Finite-State Transducer, wFST) 기반 학습인 시퀀스 분류 학습의 네 가지 알고리즘을 각각 적용하여 성능을 높이는 방법을 제안한다. 또한 기존 Cross Entropy(CE)를 사용한 학습방식과 비교하여 5 %의 상대적 word error rate(WER) 감소율을 보였다.

핵심용어: 음성인식, 트랜스포머, 시퀀스 분류 학습, 가중 유한 상태 전이기

ABSTRACT: In this paper, we adopt a transformer that shows remarkable performance in natural language processing as an acoustic model of hybrid speech recognition. The transformer acoustic model uses attention structures to process sequential data and shows high performance with low computational cost. This paper proposes a method to improve the performance of transformer AM by applying each of the four algorithms of sequence discriminative training, a weighted finite-state transducer (wFST)-based learning used in the existing DNN-HMM model. In addition, compared to the Cross Entropy (CE) learning method, sequence discriminative method shows 5 % of the relative Word Error Rate (WER).

Keywords: Speech recognition, Transformer, Sequence discriminative training, Weighted finite state transducer

PACS numbers: 43.72.Bs, 43.72.Ne

I. 서 론

하이브리드 음성인식은 초기에 음소 정보 추출을 위한 가우시안 혼합모델(Gaussian Mixture Model, GMM)과 시계열 데이터 처리에 널리 사용되는 은닉 마코프 모델(Hidden Markov Model, HMM)을 결합하여 GMM-HMM이 널리 사용되어 왔다.[1] HMM은 시계열 관측치로부터 확률을 계산하여 은닉 상태(hidden state)를 유추하는 과정을 거치면서 길이가 긴 시계열 데이터를 비교적 짧은 시계열 데이터로 바꿀 수 있고 GMM은 여러 개의 가우시안 분포를 혼합하여 특정 음성 구간의 HMM 상태를 결합 확률 분포의 형태로 나타낼 수 있다. 이를 결합한 GMM-HMM은 GMM으로 추론한 확률값과 HMM의 확률값의 결합 확률 분포를 통해 관측치로부터 음소 구간을 추정할 수 있다. 하지만 GMM은 소음에 취약하여 일부 음성

†**Corresponding author:** Joon-Hyuk Chang (jchang@hanyang.ac.kr)
Department of Electronic Engineering, Hanyang University, 222, Wangsimni-ro, Seongdong-gu, Seoul 04763, Republic of Korea.
(Tel: 82-2-2220-0355, Fax: 82-2-2291-0357)

에 대해 학습이 제대로 되지 않으면서 alignment가 잘 생성되지 않는 문제가 발생한다.

이런 문제점을 개선하기 위해서 확률 모델인 GMM을 심층 신경망(Deep Neural Network, DNN)으로 대체하여 DNN-HMM을 통해 음성인식을 하는 모델에 대한 연구가 지속적으로 이루어지고 있다.[2-4] DNN-HMM를 학습하기 위해서 GMM-HMM을 통해 음성 데이터의 프레임별 음소를 추정하는 forced alignment을 생성하여 학습하게 된다.

음성인식을 위한 인공신경망은 CE를 기반으로 각 프레임별 음소를 구분하는 방식으로 학습이 진행된다. 하지만 음성인식은 기본적으로 문장을 분류하는 문제이기 때문에 이러한 학습방식은 문장 전체에 대한 음성의 문맥적 특성을 고려하지 못한다. 이러한 문제점을 해결하기 위하여 Maximum Mutual Information(MMI),[5] boosted MMI(BMMI)[6]와 Minimum Bayes risk(MBR)[7] 등의 목적함수를 사용한 학습방식이 등장하였다. 앞서 언급한 방식들은 음소 추정으로부터 언어모델을 결합하여 발생 가능한 단어의 조합을 그래프로 나타낸 lattice라는 구조가 필요하다. 사전 준비된 lattice가 필요한 문제점을 해결하기 위해서 lattice가 필요없는 Lattice-Free MMI(LF-MMI)[8]라는 학습 방식이 등장하였다.

심층 신경망에 대한 연구가 활발히 진행되면서 다양한 구조의 심층 신경망을 음향모델로 사용하는 시도가 지속적으로 이루어지고 있다. 그 중에서 본 논문에서는 일반적인 Convolution Neural Network(CNN)와 Recurrent Neural Network(RNN)를 사용한 심층 신경망 구조가 아닌 네트워크가 데이터의 중요한 부분에 집중할 수 있도록 하는 알고리즘인 attention으로 네트워크를 구성한 트랜스포머(Transformer)[9]를 음향모델로 선택하였다. 트랜스포머는 자체적인 재귀 구조로 인한 RNN 계열 모델의 느린 학습 및 디코딩 속도를 해결 하여 상대적으로 높은 학습 속도와 인식 성능을 보인다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 DNN-HMM에 적용할 수 있는 네 가지 방법의 sequence discriminative training[10]과 관련 선행 연구 사례를 소개한다. 3장에서는 본 논문에서 수행한 우수한 성능을 가지는 트랜스포머를 기반으로 한 음향모델[11]에 기존 sequence discriminative training을 적용하는 방법에 대해 설명한다. 4장에서는 기존 CE를 사용한 학습 방식과 그래프 기반 학습 방식의 성능을 평가하며 문장 길이에 따른 성능의 변화를 비교한다. 마지막으로 5장에서 모델의 확장성을 논의하고 결론을 맺는다.

II. 관련 연구

DNN-HMM 하이브리드 음성인식에서 심층 신경망은 HMM 상태들에 대한 사후확률을 추정하는 학습을 진행한다. 이 때 주로사용하는 목적 함수는 최대우도 추정(Maximum Likelihood Estimation, MLE)이다.

$$F_{MMI} = \sum_{u=1}^{U} \log P_{\lambda}(X_u | M_{S_u}), \tag{1}$$

여기서 전체 발화 집합 U에 대해 X_u는 관측된 음성 발화, M_{S_u}는 X_u에 대응되는 인식 단위열의 집합이다.

$$y_u(s) = \frac{\exp a_u(s_r)}{\sum_s \exp(a_u(s))}. \tag{2}$$

발화 u에대한 심층 신경망의 출력값은 t시점에서의 관측값 o_u가 주어졌을 때 심층 신경망의 HMM 상태 s에 대한 softmax 활성함수를 사용하여 다음과 같이 얻어진다. 여기서 $a_u(s)$는 상태 s에 대한 출력층에서의 활성함수 출력값이다.

$$F_{CE} = -\sum_{u=1}^{U}\sum_{t=1}^{T_u} \log(y_u)(s_u). \tag{3}$$

심층 신경망의 출력값과 타겟 값을 통해 심층 신경망을 학습할 때는 Eq. (3)의 CE 손실함수를 사용하여 학습을 진행한다.

$$F_{MMI} = \sum_{u=1}^{U} \log \frac{p_{\lambda}(X_u \mid M_{S_r})^{\kappa} P(S_r)}{\sum_S p_{\lambda}(X_u \mid M_S)^{\kappa} P(S)}. \tag{4}$$

Eq. (4)는 상호정보를 최대화 하는 방식으로 네트워

크를 최적화 하는 MMI이다. 여기서 M_{S_r}는 lattice 상에서의 정답발화 S_r에 대한 HMM 상태이며 κ는 스케일 팩터이다. MMI는 사전확률이 아닌 사후확률을 직접적으로 추정하며 학습을 진행한다. 이 것이 가능 한 이유는 특정 발화에 대한 가능한 모든 단어 시퀀스를 lattice 상으로 제한함으로써 $P(S_r)$을 구할 수 있기 때문이다. 정답 발화의 단어 시퀀스에 대한 확률을 최대화하고 lattice 상의 전체 단어 시퀀스들에 대한 확률을 최소화 하여 상호 정보를 최대화하는 것이 MMI 목적함수를 사용한 인공신경망의 학습이다.

$$F_{BMMI} = \sum_{u=1}^{U} \log \frac{p_\lambda(X_u \mid M_{S_r})^\kappa P(S_r)}{\sum_s p_\lambda(X_u \mid M_S)^\kappa P(S) e^{-bA(S,S_u)}} . \quad (5)$$

Eq. (3)은 MMI 목적 함수에서 변형되어 가능한 경로들에 대한 우도를 증폭시켜 모델이 좀 더 학습하기 어려운 데이터를 잘 학습할 수 있도록 스케일 팩터 를 사용하였다. $A(S,S_u)$는 전체 단어 시퀀스에 대한 발화 u의 평균 정확도이다. CE는 프레임 단위 음소의 에러를 최소화 하고 MMI는 문장단위의 에러를 최소화한다. 반면 목적함수 MBR은 단어 단위가 아닌 세분화된 라벨에 대한 평균 에러율을 최소화하도록 설계되었다.

$$F_{MBR} = \sum_{u=1}^{U} \log \frac{\sum_S p_\lambda(X_u \mid M_{S_r})^\kappa P(S) A(S,S_r)}{\sum_{s'} p_\lambda(X_u \mid M_{S'})^\kappa P(S')} , \quad (6)$$

여기서 $A(S,S_r)$은 각 단위에 대한 정확도로서 정답 발화 S_r에 대해 각 단어 시퀀스 S에 대해 음소 라벨의 개수면 Minimum Phone Error(MPE)가 되고 HMM 상태 라벨이면 state MBR(sMBR)이 된다. 하지만 위의 lattice 기반 들은 미리 준비된 lattice가 필요하다는 단점이 존재한다. 이를 해결하고자 lattice가 필요없는 MMI(Lattice Free Maximum Mutual Inforemation, LF-MMI)가 제안되었다.

LF-MMI은 기존 MMI와 다르게 단어 기반의 언어모델이 아닌 n-gram 음소 언어모델(LM)을 사용하여 음소기반의 그래프를 생성한다. n-gram 음소 LM은 가능한 모든 단어 시퀀스들을 $denominator$ 그래프 G_{den}라는 하나의 HMM 그래프로 만든다. 따라서 LF-MMI은 Eq. (2)의 분모식을 $P(X|G_{den})$으로 치환 할 수 있다. 동일한 방식으로 Eq. (2)의 분자식을 $denominator$ 그래프 G_{den}의 정답 문장으로 이루어진 $numerater$그래프 G_{num}으로 대체할 수 있고 LF-MMI의 손실함수는 다음과 같이 정의 할 수 있다.

$$F_{LF-MMI} = \sum_{u=1}^{U} \log \frac{P(X_u|G_{num}^u)}{P(X_u|G_{den}^u)} . \quad (7)$$

DNN-HMM구조에서 심층 신경망으로 RNN이나 CNN을 사용하여 음향모델을 학습시킨 사례가 있다. 또한 트랜스포머라는 모델을 음향모델로 사용하여 학습된 HMM-GMM으로 학습 데이터의 forced alignment를 target으로 CE방식을 사용하여 학습한 방식이 존재한다. 트랜스포머는 기존에 존재하는 불필요한 부분을 마스킹하여 특정 구간을 집중할 수 있도록 하는 attention 알고리즘을 변형하여 query, key, value가 같은 self-attention 구조를 사용해 CNN이나 RNN 구조를 사용하지 않고 병렬적인 연산구조로 빠른 연산속도와 높은 성능을 보인 모델이다. 본 논문에서는 위의 다양한 sequence discriminative training 방식으로 트랜스포머 음향모델을 학습하고 결과를 비교 분석한다.

III. 트랜스포머 음향모델 학습 방법

트랜스포머 음향모델의 학습은 크게 lattice 기반과 lattice free 기반 두 가지로 구분여 진행하였다. 전자는 GMM-HMM을 mono-phone 및 tri-phone 단위로 학습하여 트랜스포머의 target이 되는 alignment와 lattice를 생성한다. 생성된 alignment로 CE 손실 함수를 사용하여 트랜스포머를 학습한다. 학습된 트랜스포머를 생성한 lattice를 기반으로 세 가지 lattice based 목적함수인 MMI, BMMI, SMBR을 사용하여 추가적인 학습을 진행한다.

LF-MMI를 사용한 트랜스포머 학습은 음소 단위 LM을 학습하는 것에서 시작한다. 학습 데이터셋의 전사 데이터로부터 발음사전을 이용하여 발음열로

전화하고 이를 통해 *denominater* 그래프를 생성하기 위한 n-gram 음소 LM을 학습하고 학습된 LM을 기반으로 그래프를 생성 및 저장한다. 마찬가지로 각 훈련 데이터 마다의 *numerater* 그래프를 생성하여 저장한다. 그래프 생성이 완료되면 트랜스포머 음향모델 학습을 위해 저장된 *denominater/numerater* 그래프를 불러와 LF-MMI 목적 함수를 사용하여 역전파를 계산해서 모델 파라미터를 학습하게 된다.

IV. 실험 및 결과

본 연구에서는 모델의 성능을 평가하기 위해 CE를 방식의 학습과 각각의 sequence discriminative training 방식의 학습을 사용한 트랜스포머 음향모델의 성능을 서로 비교한다.

4.1 실험 데이터

본 논문의 실험을 위해서 모델 학습을 위한 데이터셋은 문장 길이에 따른 각 sequence discriminative training를 적용한 트랜스포머 음향모델의 성능이 어떻게 달라지는지 확인 위하여 음성데이터의 길이가 다양한 데이터 셋이자 음성인식 분야에서 널리 쓰이는 LibriSpeech ASR corpus[12]에서 100 h 가량의 노이즈가 없는 데이터 부분집합인 train-clean-100 데이터셋을 훈련 데이터로, 노이즈가 없는 test-clean 데이터셋을 평가 데이터로 사용하였다. Train-clean-100의 음성 데이터는 짧은 문장과 긴문장으로 다양한 길이로 구성되어 있다.

또한 문장단위 학습이 짧은 문장에 대해서도 효과가 있는지 검증하기 위하여 주로 1개의 단어에서 5개의 단어로 구성되어 있는 차량용 한국어 데이터셋인 HD-100h를 사용하여 성능을 측정하였다. 해당 데이터셋은 '아웃백', '화양동 주민센터', '오승에게 전화', '경로 취소', '주공 4차 아파트' 등과 같이 주로 차량용 네비게이션에 사용되는 짧은 발화들로 구성되어 있다. HD-100h 데이터셋의 5 %, 5 %, 90 %를 각각 test, dev, train 데이터셋으로 사용하였다.

4.2 트랜스포머 음향모델

트랜스포머 음향모델은 Fig. 1과 같은 트랜스포머의 인코더를 사용하고 내부에는 multi-head attention, Feed Forward Network(FFN) 및 layer normalization으로 구성되어 있으며 end-to-end 음성인식 toolkit인 espnet[13]의 트랜스포머 인코더 구조를 변형하여 사용하였다. 연산 시간을 줄이기 위하여 두 개의 convolutional neural network(CNN) layer를 트랜스포머 subsampling layer에 추가하였다. 제안된 구조의 상세한 파라미터는 Table 1에 나타나 있다.

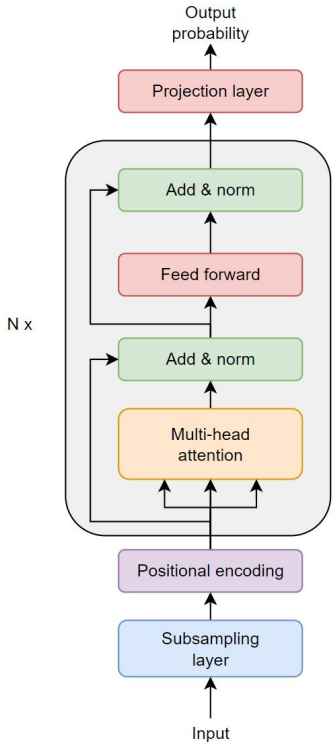

Fig. 1. (Color available online) The architecture of transformer-based AM.

Table 1. Model architecture for transformer acoustic model (AM).

	Units
Attention layer dimension	512
Encoder layer dimension	2048
Attention head	8
Encoder FFN hidden unit	2048

4.3 실험 환경

음향 모델의 전처리는 Kaldi toolkit[14]을 사용하여 진행 하였다. 전처리는 음향모델의 입력으로 사용되는 80차원의 log Mel-filter bank의 추출과 LM학습이

포함된다. 트랜스포머 모델의 경우 Pytorch[15] framework를 사용하였기 때문에 Kaldi toolkit의 모델 최적화를 위해서 Kaldi의 Python wrapper인 PyKaldi2[16]의 그래프와 lattice 생성 및 lattice-based sequence training을 사용하였다. LF-MMI학습의 경우 Kaldi의 chain 모델 loss 함수를 Pytorch에서 사용할 수 있도록 해주는 toolkit인 Pychain[17]을 사용하였다. 실험은 lattice-based sequence training과 lattice-free sequence training 두 가지로 나누어 실험을 진행한다. 트랜스포머 AM을 Librispeech와 HD-100h dataset으로 Lattice-based 방식인 MMI, BMMI, sMBR을 각각 적용하여 기존의 CE 방식의 학습과 성능을 비교한다. LF-MMI를 사용한 트랜스포머 학습은 HD-100h에 대해 실험을 진행하였으며 이는 짧은 문장에 대해서도 lattice를 사용하지 않는 sequence training이 효과적인지 증명하기 위함이다. 디코딩은 beam search 디코딩을 사용하였으며 beam size는 15, lattice beam size는 8을 사용하였고 Kaldi tookit의 rescoring을 사용하여 최종적으로 각 테스트 데이터셋에 대한 WER을 측정하였다.

4.4. 실험 결과 및 분석

본 논문에서는 CE를 사용한 학습방법과 Lattice-based 방식의 성능을 비교하고자 문장의 길이가 다양한 데이터로 구성된 Librispeech로 성능을 비교하였다. Table 2는 총 3가지 방법으로 학습한 트랜스포머 음향모델의 성능을 나타낸다. CE 기반의 학습은 test-clean에 대해 7.62 %의 WER을 나타내었고 lattice based sequence training 방식인 MMI는 7.45 %의 WER를 나타내었다. 인식 단위로 HMM state를 사용하는 sMBR의 경우에는 WER 7.32 %로 MMI보다 높은 성능으로 각각 상대 WER 감소율 2.2 % , 3.9 %를 보였다. 이를 토대로 문장 전체를 고려한 학습 방법이 frame 단위 음소 추정보다 더 효과적이라는 것을 알 수 있다.

문장이 짧을 때도 Sequence training이 효과적인지 확인하기 위하여 비교적 짧은 단어들(1~8단어)로 구성된 데이터셋인 HD-100h에 대해서도 성능을 측정한 것이 Table 3에 나타나 있다. CE를 사용한 트랜스포머 음향모델의 WER은 5.52 %이며 MMI는 5.72 %, BMMI는 5.65 %, 마지막으로 sMBR은 5.58 %가 측정

Table 2. WERs (%) on Librispeech for each lattice-based sequence training method.

	test-clean
CE	7.62
MMI	7.45
sMBR	7.32

Table 3. WERs (%) on HD-100h for each lattice-based sequence training method.

	HD-100h test
CE	5.52
MMI	5.72
BMMI	5.65
sMBR	5.58

Table 4. WERs (%) on 7 different word count sections of Librispeech.

word count	CE	MMI
1 ~ 5	10.02	10.51
5 ~ 10	9.22	9.39
10 ~ 20	7.80	7.66
20 ~ 40	6.88	6.87
40 ~ 60	7.05	6.92
60 ~ 80	7.89	7.75
80 ~	10.08	10.08

되었다. Sequence training방식의 학습이 기존 CE 방식보다 짧은 발화로 구성된 HD-100h에서 더 낮은 성능 보였다.

Table 4는 실제로 짧은 문장에서 sequence training 방식이 효과적인지 못한지 확인하기 위하여 Table 2에서 사용한 Librispeech 데이터셋에 대해 문장 단어 개수별 총 7개의 구간으로 구분하여 MMI 방식과 CE 방식의 성능을 측정한 것을 나타낸다. Lattice-based 방식인 MMI가 단어 1 ~ 5개, 5 ~ 10개로 이루어진 문장들에 대해서 CE보다 WER 수치가 더 높은 것을 확인하였다. 반면 10 ~ 80개의 단어로 구성된 문장들에 대해서는 MMI의 WER 수치가 CE보다 낮기 때문에 lattice-based 방식이 짧은 문장에서 효과적이지 못하다는 것을 보여준다.

Table 5의 결과를 보면 Lattice를 사용하지 않은 sequence training인 LF-MMI가 HD-100h 데이터셋에서 WER 5.2 %로 기존 CE 방식보다 5 %의 상대 WER

Table 5. WERs (%) on HD-100h for LF-MMI sequence training method.[6]

	HD-100h test
CE	5.52
LF-MMI	5.20

Table 6. WERs (%) on Librispeech for CE, sMBR and LF-MMI.

	test-clean
CE	7.62
sMBR	7.32
LF-MMI	7.20

감소를 보였다. Lattice를 사용하지 않는 방식의 LF-MMI가 짧은 문장들에 대해서도 효과적이라는 것을 알 수 있다.

마지막으로 Table 6은 Librispeech 데이터셋에 대한 각 sequence training의 성능을 비교한 것으로 LF-MMI가 기존 CE 방식보다 약 5.5 %, sMBR에 대해서는 약 1.7 %의 상대 WER 감소율을 보였다. 이는 문장의 길이와 상관없이 LF-MMI가 가장 성능이 뛰어나다는 것을 보여준다.

V. 결 론

본 연구에서는 기존 음성인식에서 뛰어난 성능을 보이는 트랜스포머 인코더 구조를 변형하여 하이브리드 음성인식에서의 음향모델로 사용하였다. 심층 신경망 음향모델의 학습은 음성의 프레임마다 forced alignment를 정답으로 활용하여 CE 손실함수로 학습한다. 이 학습방식은 문장 전체를 고려하지 못하는 단점이 있다. 이러한 단점을 극복하기 위하여 sequence discriminative training을 트랜스포머 음향모델에 적용하여 그래프 기반의 학습을 적용하였다.

Lattice-based 방식의 sequence training은 긴 문장에 대해 기존의 CE 방식보다 약 3.9 %의 성능 향상을 보였으나 짧은 문장에 대해서 성능이 떨어지는 단점이 존재하였다. 하지만 lattice를 사용하지 않는 LF-MMI는 4-gram 음소 언어모델을 기반의 wFST를 사용하기 때문에 짧은 문장의 데이터셋에 대해 기존 CE 방식보다 약 5 %의 상대적인 성능 개선을 보였다.

감사의 글

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2021-0-00456, 원격 다자간 영상회의에서의 음성 품질 고도화 기술개발).

References

1. B. Juang and L. Rabiner, “Hidden Markov models for speech recognition,” Technometrics, **33**, 251-272 (1991).
2. A. Senior, H. Sak, and I. Shafran, “Context dependent phone models for LSTM RNN acoustic modelling,” Proc. IEEE ICASSP, 4585-4589 (2015).
3. J. Li, V. Lavrukhin, B. Ginsburg, and R. Leary, “Jasper: An end-to-end convolutional neural acoustic model,” arXiv preprint arXiv:1904.03288 (2019).
4. K. Chen and Q. Huo, “Training deep bidirectional LSTM acoustic model for LVCSR by a context-sensitive-chunk BPTT approach,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24 (2016).
5. L. Bahl, P. Brown, P. Souza, and R. Mercer, “Maximum mutual information estimation of hidden Markov model parameters for speech recognition,” Proc. ICASSP, 49-52 (1986).
6. D. Povey, D. Kanevsky, B. Kingsbury, B. Ranabhadran, G. Saon, and K. Visweswariah, “Boosted MMI for model and feature-space discriminative training,” Proc. IEEE ICASSP, 4057-4060 (2008).
7. M. Gibson and T. Hain, “Hypothesis spaces for minimum Bayes risk training in large vocabulary speech recognition,” Proc. Interspeech, 2406-2409 (2006).
8. D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, and V. Manohar, “Purely sequence-trained neural networks for ASR based on lattice-free MMI,” Proc. Interspeech, 2751-2755 (2016).
9. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in neural information processing systems, 30 (2017).
10. K. Vesely, A. Ghoshal, L. Burget, and D. Povey, “Sequence-discriminative training of deep neural networks,” Proc. Interspeech, 2345-2349 (2013).
11. Y. Wang, A. mohamed, D. Le, C. Liu, and A. Xiao, “Transformer-based acoustic modeling for hybrid speech recognition,” Proc. IEEE ICASSP, 6874-6878 (2020).
12. V. Panayotov, G. Chen, D. Povey, and S.Khudanpur, “Librispeech: an asr corpus based on public domain

audio books," Proc. IEEE ICASSP, 5206-5210 (2015).
13. S. Watanabe, T. Hori, S. karita, and T. Hayashi, "Espnet: End-to-end speech processing toolkit," arXiv preprint arXiv:1804.00015 (2018).
14. 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).
15. A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, and Z. Vito, "Pytorch: An imperative style, high-performance deep learning library," Advances in neural information processing systems, 32 (2019).
16. L. Lu, X. Xiao, Z. Chen, and Y. Gong, "Pykaldi2: Yet another speech toolkit based on kaldi and pytorch," arXiv preprint arXiv:1907.05955 (2019).
17. Y. Shao and Y. Wang, "Pychain: A fully parallelized pytorch implementation of lf-mmi for end-to-end asr," arXiv preprint arXiv:2005.09824 (2020).

저자 약력

▸ 이 채 원 (Chae-Won Lee)

2020년 8월 : 한양대학교 융합전자공학부 학사
2020년 9월 ~ 현재 : 한양대학교 융합전자공학과 석박사통합과정 재학 중

▸ 장 준 혁 (Joon-Hyuk Chang)

2004년 2월 : 서울대학교 전기컴퓨터공학부 박사
2000년 3월 ~ 2005년4월 : (주)넷더스 연구소장
2004년 5월 ~ 2005년4월 : 캘리포니아 주립대학 산타바바라(UCSB) 박사후 연구원
2005년 5월 ~ 2005년8월 : 한국과학기술연구원(KIST) 연구원
2005년 9월 ~ 2011년 2월 : 인하대학교 전자공학부 조교수
2011년 3월 ~ 2017년 3월 : 한양대학교 융합전자공학부 부교수
2017년 3월 ~ 현재 : 한양대학교 융합전자공학부 정교수