한국음향학회지 제40권 제5호 pp. 488~495 (2021)
The Journal of the Acoustical Society of Korea Vol.40, No.5 (2021)
https://doi.org/10.7776/ASK.2021.40.5.488

pISSN : 1225-4428
eISSN : 2287-3775

콘포머 기반 한국어 음성인식

A Korean speech recognition based on conformer

구명완[1†]

(Myoung-Wan Koo[1†])

[1]서강대학교 컴퓨터공학과

(Received August 2, 2021; accepted September 14, 2021)

초 록: 본 논문에서는 콘포머 기반 한국어 음성인식 시스템을 제안한다. 콘포머는 트랜스포머 모델에 콘볼루션신경망(Convolution Neural Network, CNN) 기능을 보강한 구조이며 광역 정보를 잘 표현할 수 있는 트랜스포머와 지역 정보를 잘 표현할 수 있는 CNN을 결합한 신경망이다. 음성인식 기본 시스템으로 트랜스포모에 기반한 음성인식시스템을 개발하였으며 언어모델로는 Long Short-Term Memory(LSTM)을 사용하였다. 콘포머 기반 음성인식시스템은 트랜스포머 대신에 콘포머를 사용하였고 언어모델로는 트랜스포머를 이용하였다. 성능 평가를 위해 AI-hub에 있는 Electronics and Telecommunications Research Institute(ETRI) 음성코퍼스를 활용하였으며 트랜스포머 기반 음성인식 시스템은 오인식률이 11.8 %이 되었으며 콘포머 기반 음성인식시스템은 오인식률이 5.7 %가 되었다. AI-hub에 있는 다른 영역의 NHN다이퀘스트 음성 코퍼스를 추가해도 유사한 성능이 유지가 되어 제안된 콘포머 음성인식시스템의 유효성을 입증하였다.

핵심용어: 음성인식, 딥 러닝, 콘포머, 트랜스포머

ABSTRACT: We propose a speech recognition system based on conformer. Conformer is known to be convolution-augmented transformer, which combines transfer model for capturing global information with Convolution Neural Network (CNN) for exploiting local feature effectively. The baseline system is developed to be a transfer-based speech recognition using Long Short-Term Memory (LSTM)-based language model. The proposed system is a system which uses conformer instead of transformer with transformer-based language model. When Electronics and Telecommunications Research Institute (ETRI) speech corpus in AI-Hub is used for our evaluation, the proposed system yields 5.7 % of Character Error Rate (CER) while the baseline system results in 11.8 % of CER. Even though speech corpus is extended into other domain of AI-hub such as NHNdiguest speech corpus, the proposed system makes a robust performance for two domains. Throughout those experiments, we can prove a validation of the proposed system.

Keywords: Speech recognition, Deep learning, Conformer, Transformer

PACS numbers: 43.60.Fg, 43.72.Dv

I. 서 론

4차 산업혁명에 인공지능이 중요한 역할을 하게 된 최초의 연구는 음성인식 분야에서 신경망을 사용하게 된 것이다. 신경망을 음성인식 분야에 적용한 결과 10년간 정체가 되어온 음성인식 성능이 비약적으로 향상이 되었다. 즉 Hidden Markov Model(HMM)을 활용한 통계적 방식에서 음향특징을 나타내는 음향 확률분포에 다중 층 퍼셉트론(multi-layer perceptron) 혹은 Long Short-Term Memory(LSTM)을 사용한 결과 동일한 음성 코퍼스를 활용할 경우 약 16 %의 음성인식률이 향상이 되었다.[1] 이 후 많은 연구자들이

†**Corresponding author:** Myoung-Wan Koo (mwkoo@sogang.ac.kr)
Department of Computer Science and Engineer, Sogang University, 35 Baekbum-ro, Mapo-gu, Seoul 04107, Republic of Korea
(Tel: 82-2-705-8935, Fax: 82-2-526-5909)

신경망을 활용하는 방안을 제안하였으며 대표적인 모듈러 방식인 HMM기반 음성인식 시스템 대신에 음성을 입력하면 인식 결과를 한 번에 제공할 수 있는 종단형 음성인식 시스템이 최근에 개발이 되고 있다.[2] 종단형 음성인식 시스템은 기계 변역에 사용되는 신경망에 기반하며 음성의 특징을 이해하는 인코더와 그것에 기반하여 문자를 출력하는 디코더로 나누어진다.[3]

종단형 음성인식시스템의 대표적인 방식으로 Connectionist Temploral Classification(CTC)가 있는데 이 것은 프레임 단위로 처리가 가능하며 Bidirectional LSTM (BLSTM)을 사용하여 특징을 추출하고 매 프레임 단위로 브랭크를 포함한 음소단위로 출력을 하게 되어 속도가 빠르고 온라인 처리가 가능하다는 장점이 있지만 출력 음소 사이의 연관성이 없다는 조건으로 성능이 떨어지는 단점이 있다.[4] 기계 번역에 사용이 되고 있는 어텐션 기반 인코더 디코더를 종단형 음성인식 시스템에 그대로 사용하는 방안도 제안되었다.[5] 이 방식은 인코더에는 음성특징을 디코더는 언어모델을 이용할 수 있으며 디코더에 조건 독립 가정이 필요 없으므로 성능이 우수하다는 장점이 있으나 속도가 느리고 복잡하다는 단점도 있다. Recurrent Neural Network(RNN)-transducer는 CTC를 확장한 개념으로 입력 특징으로 BLSTM을 사용하고 출력을 위해서 RNN 출력을 다시 입력으로 사용하는 자기회귀 RNN을 사용한 개념으로 성능도 좋으며 온라인이 가능한 구조이나 구현이 복잡하고 느린 단점이 있다.[6]

트랜스포머는 자기 집중기능이 있어서 LSTM보다 인코더에서 음성 특징을 더 잘 추출할 수 있으며 디코더에 사용할 경우에도 출력을 위한 자기회귀 기능이 없으므로 LSTM 보다 성능이 우수하다는 장점이 있어서 LSTM 대신에 많이 사용되고 있다.[7] Convolutional Neural Networks(CNN) 알고리즘은 영상분야에서는 많이 사용이 되고 있으나 음성분야에서는 음성 코딩 및 합성 분야에서 주로 사용이 되어 왔으나 음성인식 분야에서는 사용이 된 예가 매우 적다.[8]

본 논문에서는 CNN과 트랜스포머 방식을 결합한 콘포머 방식을 사용한 음성인식 시스템을 제안하고 트랜스포머 방식과 성능을 비교한다. 2절에서는 트랜스포머와 콘포머에 대해서 비교 설명하고 3절에서는 콘포머 음성인식 시스템의 훈련 및 인식 알고리즘을 소개한다. 4절에서는 한국어 음성코퍼스를 사용한 성능 평가를 트랜스포머와 콘포머를 비교한다. 5절에서는 결론을 맺는다.

II. 트랜스포머와 콘포머

최근에 사용이 되고 있는 딥러닝 알고리즘으로 트랜스포머와 콘포머가 있는 데 각각에 대해서 소개를 한다.

2.1 트랜스포머

트랜스포머는 기계번역에 최초로 사용이 되었으며 Fig. 1과 같이 인코더와 디코더로 구성이 되어 있다.[9]

인코더는 입력의 특징을 잘 표현하도록 하고 디코더는 출력에 맞도록 입력과의 연결고리를 잘 표현할 수 있도록 훈련이 될 수 있는 것이 트랜스포머의 특징이다. 실제로 입력을 영어로 출력을 독일어로 정한 기계 번역에 대한 실험을 한 결과 기존 방식에 비해서 가장 우수한 기계 번역의 결과가 나타났다. 음성인식

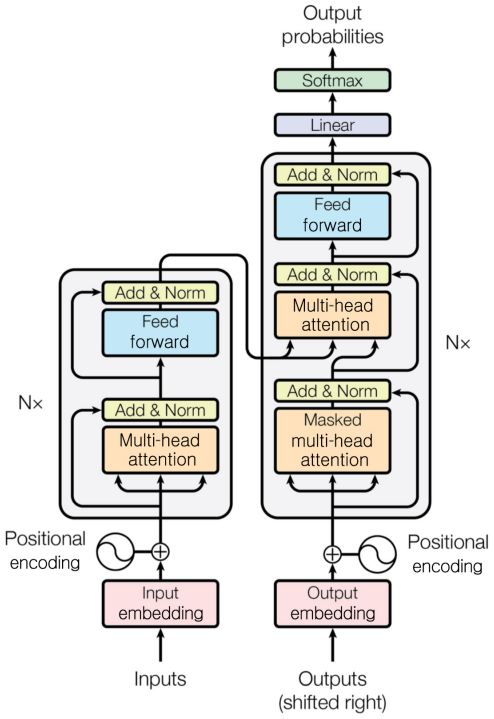

Fig. 1. (Color available online) The model architecture of transformer.

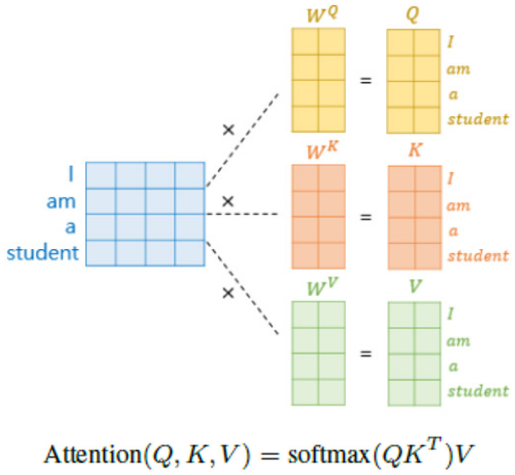

Fig. 2. (Color available online) Self-attention in sentence "I am a student".

의 경우에는 입력이 음성이 되고 출력이 문자가 되는 것이고 음성합성의 경우는 반대가 되는 것이다.

트랜스포머가 기존의 알고리즘에 비해서 우수한 성능을 제공할 수 있는 이유는 인코더, 디코더에 있는 멀티헤드 어텐션 기능이다. 멀티헤드 어텐션은 자기 집중기능을 여러 개를 활용하는 것으로 Fig. 2에 그 개념이 나와 있다. 예를 들면 "I am a student" 라는 문장에 대한 자기 집중 기능을 사용하면 문장을 구성하고 있는 각 단어에 대해서 연관이 되는 단어 사이의 중요도를 Q(query), K(key), V(value)로 표현한 것이다. 즉 Q의 해당하는 값이, K, V 각각의 값과 어떻게 연관성이 있는지를 잘 표현해 주는 무게벡터를 목표치에 도달하도록 구하는 방식이다. 기존의 CNN, LSTM 알고리즘은 입력과 출력 사이의 무게벡터만 구하였다면 트랜스포머 방식은 입력 사이의 연관도도 같이 본다는 것이 차이점이다.

2.2 콘포머

콘포머는 Fig. 1에서 트랜스포머 인코더 대신에 콘포머 인코더가 사용이 되며 디코더는 트랜스포머를 사용한다. 콘포머 인코더의 구성은 Fig. 3에 나나타나 있듯이 트랜스포머에 있는 Feed Forward Network (FFN)가 두 개로 나누어지고 멀티헤드 언텐션 다음에 콘볼루션 모듈이 추가가 되었다. Fig. 3에서 트랜스포머 2개에서 FFN을 2개로 분리해서 마카론처럼 어텐션 기능을 감싼 것인데 구조도는 Fig. 4에 표시

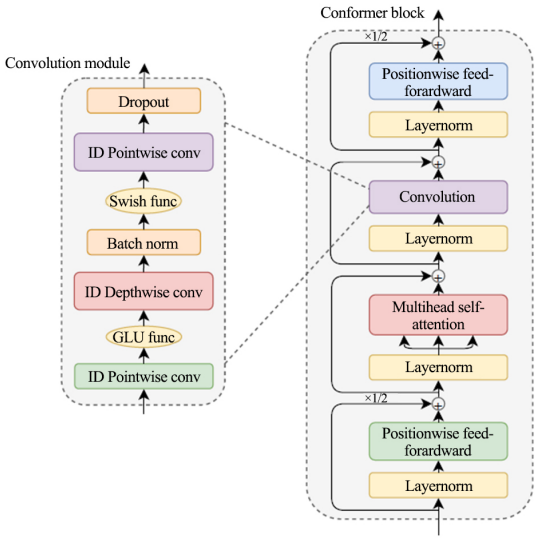

Fig. 3. (Color available online) Overview of conformer block.

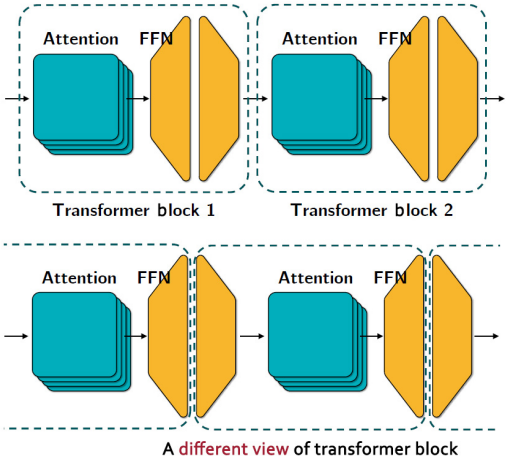

Fig. 4. (Color available online) Different view of transformer block.

되어 있다. 콘포머에서 멀티헤드 어텐션 기능은 수식 Eq. (1)에 표시가 되어 있다.

$$MHSA(Q,K,V) = Concat(head_1,...,head_H)\,W^0$$
$$head_1 = Attention\,(Q_h,K_h,V_h). \tag{1}$$

FFN은 전통적인 정류선형 유니트(Rectified Linear Unit, ReLU) 동적 함수를 갖고 있는 신경망이며 상세한 수식은 Eq. (2)에 표현이 되어 있다.

$$FFN(X) = W_2\,ReLU(W_1X+b_1)+b_2. \tag{2}$$

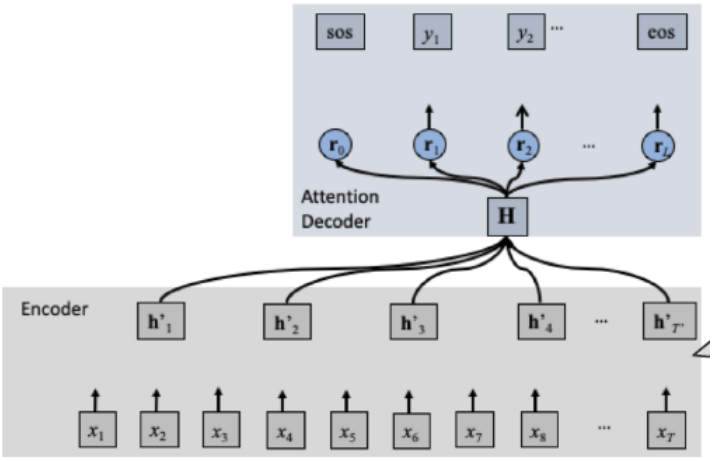

Fig. 5. (Color available online) Block diagram of speech recognition system based on conformer.

Fig. 3의 콘볼루션 모듈은 1차원 콘볼류션 층과 Gated Linear Units(GLU) 활동 함수를 사용한다.[10] 1차원 포인트와이즈 콘볼루션 층에서는 입력의 컨불류션을 2개 사용하여 입력 채널을 두 배로 늘리며 하나의 채널이 게이트의 역할을 하도록 하여 1채널의 결과를 구하고 그 결과를 소프트맥스한다. 다음에는 배치 정규화를 통해서 입력 값을 정규화하여 훈련이 잘되도록 하며 다음에는 스위쉬 함수를 사용하여 1차원 포인트와이즈 콘볼루션을 다시 수행한다. 스위쉬 함수는 Relu 함수 보다 우수한 훈련 결과를 나타내고 있으며,[11] 1차원 컨불류션은 마카론 구조와 동일한 효과를 얻기 위한 것이다. 마지막으로 사용되는 드롭아웃트는 훈련시 과적화 현상을 방지하기 위한 것이다.

III. 콘포머 기반 한국어 음성인식

본 논문에서 제안하고 있는 콘포머 기반 음성인식 시스템은 인코더를 콘포머를 사용하고 디코더는 어텐션 기반 트랜스퍼 디코더를 병행한 구조이다. 또한 트랜스퍼 기반 언어모델도 사용하였다. Fig. 5는 인코더로 콘포머를 사용하고 출력으로 트랜스포머를 사용하고 있는 콘포머 기반 음성인식 시스템 구조도를 나타내고 있으며 Table 1에 각 레이어별 상세 파라미터가 나타나 있다. 제안된 구조는 트랜스포머 구조에서 인코더로 콘포머를 변경하였으며 디코더는 기존의 트랜스포머 구조를 그대로 사용하였다. 그 이유는 콘포머 특징이 음성으로부터 지역 및 글로벌 특징을 추출하는데 특화가 되어 있으므로 인코더만 변경하는 것이 바람직하기 때문이다.

Table 1. Neural network architecture for speech recognition system based on conformer.

	Units
Mel	80
Spec augmentation	freq. mask [0,30] time mask [0,40]
embedding(vocab,dim)	40 msec, (5000, 512)
Encoder head	8
Encoder FFN hidden unit	2048
Encoder layer	12
Decoder head	8
Decoder FFN hidden unit	2048
Decoder layer	6

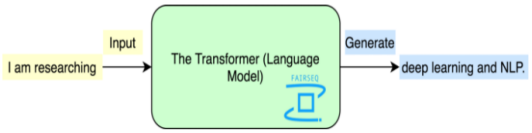

Fig. 6. (Color available online) Block diagram of transformer-based language modeling.

본 논문에서는 트랜스포머 기반 언어 모델를 제안하며 Fig. 6에 개념도가 그려져 있다. LSTM에 기반한 언어 모델과의 차이점은 LSTM 기반 언어 모델은 토큰(단어) 단위로 회기적으로 출력이 된다면 트랜스포머 기반 언어 모델은 입력 과 출력이 토큰 집합 단위로 처리가 된다는 것이다. 즉 Fig. 6에서 입력이 "I am researching"가 되면 트랜스포머 출력은 "deep

Table 2. Neural network architecture for language modeling using transfer network.

	Number
Embedding(vocab, dim)	(5000, 128)
Attention unit	512
Head	8
FFN hidden unit	2048
Layer	16
Memory	214.8 MB

Table 3. Neural network architecture for speech recognition system based on transformer.

	Units
Mel	80
Spec augmentation	freq. mask [0,30] time mask [0,40]
embedding (vocab,dim)	40 msec, (5000, 512)
Encoder head	8
Encoder FFN hidden unit	2048
Encoder layer	12
Decoder head	8
Decoder FFN hidden unit	2048
Decoder layer	6

learning and NLP"가 된다는 것이다. 그러므로 언어모델을 위한 트랜스포머는 디코더 한 개만 있으면 된다. 언어모델로 트랜스포머를 사용한 이유는 기존에 많이 사용되고 있는 LSTM보다 성능이 우수하고 실시간 처리가 가능하다는 장점이 있기 때문이다. 언어모델의 신경망 구조도에 대한 상세 스펙은 Table 2에 있다.

베이스라인으로 사용이 되는 트랜스포머 기반 한국어 음성인식 시스템은 Fig. 5의 인코더로 트랜스포머를 사용하는 것이고 Table 3에 인코더 및 디코더의 상세한 레이어 별 파라미터 스펙을 표시하였다. 비교 실험을 위하여 모든 파라미터는 같으며 단지 콘포머 기반 음성인식 시스템은 Fig. 3에서 콘볼루션 모듈만 추가가 된 것이다.

IV. 실험 및 결과

4.1 데이터 베이스

본 논문의 실험을 위해서 AI hub(https://aihub.or.kr) 에서 2018년도에 Electronics and Telecommunications Research Institute(ETRI)가 구축한 1000 h 분량의 한국어 음성과 2021년도에 NHN 다이퀘스트가 구축한 한국어 자유대화 음성 코퍼스를 활용하였다. 각각 음성 코퍼스는 훈련, 검증, 실험을 위해서 8:1:1로 나누어서 실험을 수행하였다. Table 4에서는 NHN 다이퀘스트가 구축한 음성 코퍼스를 실험에 적합하게 나눈 데이터베이스를 자세히 기술하였다. Table 5에는 ETRI가 구축한 음성 코퍼스에 대한 정보가 나타나 있다. ETRI 코퍼스는 안부 일상대화, 날씨, 쇼핑, 취미 등에 관해서 자유롭게 대화한 것을 전사한 것이다. 훈련, 검증, 실험 데이터 셋의 비율은 8:1:1로 나누었다.

Table 4. Data set for NHNdiquest corpus.

	Speaker	Sentence	Train	Eval.	Test
Adult	55	77,270	66,876	7121	3273
Children	55	68,846	58,612	6066	4168
Senior	53	56,744	48,240	5836	2668
Loan word	55	85,028	73,187	8146	3695
Total	218	287,888	246,915	27,169	13,804

Table 5. Data set for ETRI corpus.

	Speaker	Sentence	Train	Eval.	Test
Total	1000	615,368	585,715	14,823	14,830

언어 모델 훈련을 위한 언어 코퍼스는 Tables 4와 5의 훈련에 사용하는 언어 코퍼스 832,630 문장 이외에도 인터넷에서 구한 언어 코퍼스 926,547 문장을 추가해서 17만 문장으로 훈련을 하였다.

4.2 성능 평가

성능평가를 위한 기본 시스템으로 Table 3에 기반하는 트랜스포머 기반 음성인식기를 개발 하였으며 이 때 사용한 언어모델은 LSTM에 기반한 언어모델을 사용하였다. Table 6에는 LSTM에 기반한 언어모델의 상세 스펙을 나타내었다.

ETRI 코퍼스에 대한 성능 결과표는 Table 7에 표시되어 있다. LSTM에 기반한 언어모델의 성능은 117 perplexity(ppl)이 되고 음절 오인식율(Character Error Rate, CER)은 11.8 %가 된다. 이 성능은 다른 논문에서 보고된 최고 성능 10.31 %와 유사하다.[12] 단어 오인식율 대신에 음절 오인식율을 사용한 이유는 본

Table 6. Neural network architecture for language modeling using LSTM network.

	Number
Embedding (vocab, dim)	(5000, 2048)
Hidden unit	2048
Layer	4
Memory	619 MB

Table 7. Performance of transformer with ETRI corpus.

Ppl	Sub	Del	Ins	CER
117	4.9 %	5.1 %	1.8 %	11.8 %

Table 8. Performance of conformer with ETRI corpus.

Ppl	Sub	Del	Ins	CER
120	2.5 %	1.9 %	1.3 %	5.7 %

Table 9. Performance of conformer with NHNdiquest corpus.

Ppl	Sub	Del	Ins	CER
624	7.8 %	9.3 %	0.8 %	17.9 %

음성인식 시스템이 센텐스피스를 사용하여 단어 대신에 5,000개의 토큰으로 나누었기 때문에 토큰의 개수에 따라 단어 오인식율의 성능이 달라지므로 다는 논문의 성능과 비교하기가 어렵기 때문이다.

동일한 코퍼스를 활용하여 본 논문에서 제안한 콘포머 기반 음성인식 성능을 비교하였다. 사용한 콘포머 파라미터는 Table 1을 활용하였으며 언어모델은 트랜스포머을 활용하여 Table 2의 파라미터를 사용하였다. 성능 결과는 Table 8에 나타나 있다. 트랜스포머에 기반한 언어 모델의 성능은 120 ppl이 나와서 LSTM에 기반한 언어모델의 성능 117 ppl과 유사하게 나왔다. 그러나 사용한 메모리의 양은 Table 6의 618 MB에서 Table 2의 214 MB로 약 35 % 정도만 사용하였다. 반면 콘포머를 활용한 음성인식 성능은 트랜스포머에 비해서 6.1 %가 향상이 되었다.

NHN다이퀘스트 코퍼스를 활용하여 콘포머 기반 음성인식 시스템에 대한 성능 평가를 하였다. 언어모델은 트랜스포머에 기반한 방식을 사용하였으며 Table 9에 결과가 나타나 있다. 언어모델의 성능이 매우 낮은 624 ppl 로 나온 원인을 분석한 결과 훈련에 없는 외래어가 테스트 평가 셋에 많이 존재하여 언어모델이 제대로 훈련이 되지 않았다는 것을 발견하였으며 그로 인해 전반적인 음성인식 성능이 좋지 않았다.

Table 10. Performance of conformer with (ETRI + NHNdiquest) corpus.

Corpus	Ppl	Sub	Del	Ins	CER
ETRI	120.2	2.5 %	2.1 %	1.3 %	5.9 %
NHN Diguest	44.9	1.8 %	1.1 %	0.6 %	3.5 %

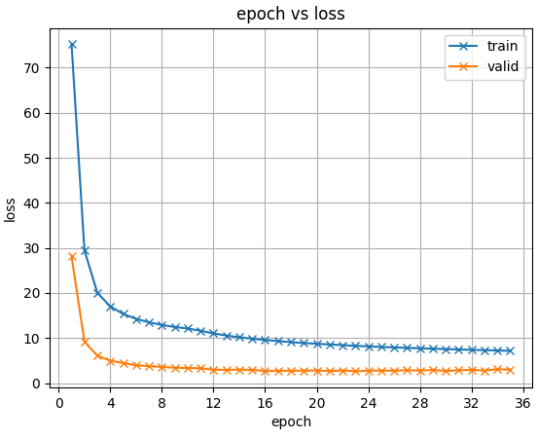

Fig. 7. (Color available online) Loss with regard to epochs during training.

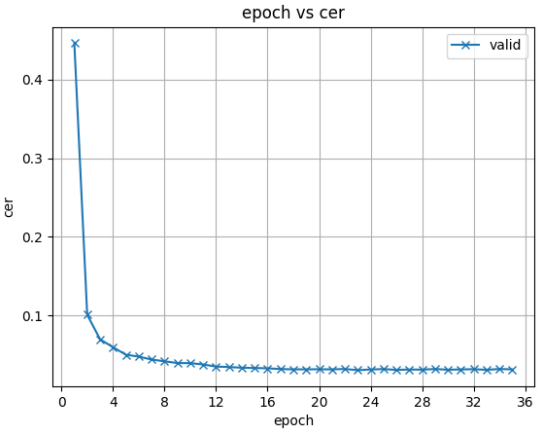

Fig. 8. (Color available online) CER with regard to epochs during training.

마지막으로 ETRI 코퍼스와 NHN다이퀘스트 코퍼스를 훈련, 검증, 평가의 단위로 두 개의 코퍼스를 합해서 훈련, 검증, 평가를 수행하였다. 그 결과는 Table 10에 표시가 되었다. 또한 훈련 과정에서의 손실 값과 오인식률(Character Error Rate, CER)을 훈련 에포크단위로 표시한 것이 Figs. 7과 8에 제시 되었다. Fig.

7에 표시된 loss는 Mean Sequare Error(MSE) loss이다. 훈련은 40 에포크에 도달하기 전 오인식률이 가장 작은 모델을 구하도록 하였다.

Table 10의 결과를 보면 다양한 음성 코퍼스와 언어 모델을 위한 다양한 언어 표현이 중요하며 그런한 코퍼스를 합쳐서 훈련을 하면 부족한 훈련 코퍼스 양 혹은 균형이 잡혀져 있는 못한 훈련 코퍼스가 있는 태스크(NHN 다이퀘스트)에서도 우수한 성능을 보이며 기존의 태스크(ETRI 분야)에서도 성능이 유지가 된다는 것을 알 수 있다. 또한 Figs. 7과 8을 보면 훈련 과정에서 약 20 에포크가 넘으면 성능이 안정적으로 유지가 된다는 것을 알 수 있다.

본 논문의 최종 결과 성능에 따르면 콘포머 기반 음성인식 시스템이 매우 성능이 우수하며 특히 ETRI 코퍼스를 활용한 한국어 관련 음성인식 연구 중에서 최고 수준의 결과가 제시되었다.

V. 결 론

본 논문에서는 콘포머 기반 한국어 음성인식 시스템을 제안하였다. 코포머는 트랜스포머의 기본 구조에서 콘볼류션 모델을 추가하고 트랜스포머 구조에서 FFN을 두 개로 나눈 마카론 구조를 갖고 있다. 음성인식 기본 시스템으로 트랜스포머에 기반한 음성인식 시스템을 개발하였으며 LSTM에 기반한 언어모델을 활용하였을 경우에 ETRI 음성코퍼스에 대해서 11.8 %의 오인식률이 나왔다. 이 성능은 유사한 연구를 수행한 한국어 최고 성능과 유사한 결과였다. 언어모델을 LSTM에서 트랜스포머 기반으로 변경하였을 경우 문법 복잡도는 유사하였으나 필요한 메모리 양이 1/3로 줄었다. 트랜스포머 대신에 콘포머를 사용하면 오인식률이 5.7 %로 줄어듬을 확인하였다.

NHN다이퀘스트가 개발한 음성코퍼스를 이용하면 실험 평가 셋의 불균형으로 오인식률이 17.9 %로 성능저하가 발생하였지만, ETRI 코퍼스와 합쳐서 훈련하면 3.5 %의 오인식률로 성능이 향상되었다. 또한 ETRI 코퍼스에 대해서도 5.9 %의 오인식률이 되어 멀티 도메인에서도 콘포머 기반 음성인식 시스템이 안정적으로 동작한다는 것을 보였다.

향후 연구로서, 콘포머 기반 한국어 음성인식 시스템이 실시간으로 동작할 수 있는 시스템의 개발이 필요하며 다양한 영역에서도 안정적인 성능이 유지될 수 있는 모델의 개발이 필요하다.

감사의 글

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No. 2021-0-00016, 엣지 컴퓨팅 기반 음성 위협 인지 기술 개발).

References

1. S. Zhao, X. Xiao, Z. Zhang, T. N. T. Nguyen, X. Zhong, B. Ren, L. Wang, D. L. Jones, E. S. Chng, and H. Li, “Robust speech recognition using beamforming with adaptive microphone gains and multichannel noise reductio,” Proc. 2015 IEEE ASRU. 460-467 (2015).
2. Y. Tachioka, T. Narita, L. Miura, T. Uramoto, N. Monta, S. Uenohara, K. Furuya, S. Wanatanabe, and J. Le Roux, “Coupled initialization of multi-channel non-negative matrix factorization based on spatial and spectral information,” Proc. Interspeech, 2461-2465 (2017).
3. D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos, E. Elsen, J. Engel, L. Fan, C. Fougner, T. Han, A. Hannun, B. Jun, P. LeGresley, L. Lin, S. Narang, A. Ng, S. Ozair, R. Prenger, J. Raiman, S. Satheesh, D. Seetapun, S. Sengupta, Y. Wang, Z. Wang, C. Wang, B. Xiao, D. Yogatama, J. Zhan, and Z. Zhu, “Deep speech 2: end-to-end speech recognition in English and Mandarim,” arXiv:1512.02595v1 (2015).
4. A. Graves and N. Jaitly, “Towards end-to-end speech recognition with recurrent neural networks,” Proc. ICML. 1764-1772 (2014).
5. W. Chan, N. Jaitly, Q. Le, and O. Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” Proc. ICASSP. 4960-4964 (2016).
6. A. Graves, A. r. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” arXiv:1303.5778 (2013).
7. L. Dong, S. Xu, and B. Xu, “Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition,” Proc. ICASSP. 5884-5888 (2018).
8. A. Oord, S. Dieleman, H. Zen, K. Simonyan, O.

Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “Wavenet: a generative model for raw audio,” arXiv:1609.03499 (2016).
9. N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” Proc. NIPS 1-11 (2017).
10. Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, “Language modeling with gated convolutional networks,” arXiv:1612.08083v3 (2017).
11. P. Ramachandran, B. Zoph, and Q. V. Le, “ Swish: A self-gated activation function,” arXiv:1710.05941v1 (2017).
12. S. Kim, S. Bae, and C. Won, “Open-source toolkit for end-to-end Korean speech recognition,” Software Impacts, **7**, 1-4 (2021).

저자 약력

▸ 구 명 완 (Myoung-Wan Koo)

1982년 2월 : 연세대학교 전자공학과 학사
1985년 2월 : 한국과학기술원 전기및전자공학과 석사
1991년 2월 : 한국과학기술원 전기및전자공학과 박사
1985년 4월 ~ 2012년 7월 : KT 상무보
1996년 11월 ~ 1997년 12월 : 미국 벨연구소 방문연구원(연구재단 post-doc fellowship)
2012년 8월 ~ 현재 : 서강대학교 컴퓨터공학과 교수