pISSN 2005-8063
eISSN 2586-5854
2024. 9. 30.
Vol.16 No.3
pp. 79-86

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2024.16.3.079

A comparative study on the performance of Transformer-based models for Korean speech recognition*

Changhan Oh[1] · Minseo Kim[2] · Kiyoung Park[1,**] · Hwajeon Song[1]

[1]*Integrated Intelligence Research Section, Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea*
[2]*Department of English Linguistics & Language Technology (ELLT), Hankuk University of Foreign Studies, Seoul, Korea*

Abstract

Transformer models have shown remarkable performance in extracting meaningful information from sequential input data such as text and images, and are gaining attention as end-to-end models for speech recognition. This study compared the performances of the Transformer speech recognition model and its enhanced versions, the Conformer and E-Branchformer, when applied to Korean speech recognition. Using Korean speech data from AIHub, we prepared a training set of approximately 7,500 hours and evaluated the models using the ESPnet toolkit. Additionally, we compared syllables and subwords as recognition units and analyzed the performance differences with changes in the number of tokens using Byte Pair Encoding. The results showed that the E-Branchformer achieved the best performance in Korean speech recognition and Conformer outperformed Transformer but degraded in performance for long utterances owing to cross-attention alignment errors. We aimed to determine the optimal settings by analyzing the performance changes with subword token adjustments. This study comprehensively evaluated model accuracy and processing speed to maximize the efficiency of Korean speech recognition. This is expected to contribute to the training of large-scale Korean speech recognition models and improve Conformer recognition errors. Future research should include additional experiments with diverse Korean speech datasets and enhance the recognition performance through structural improvements in the Conformer.

Keywords: deep learning, machine learning, speech recognition

1. 서론

텍스트, 영상 등과 같이 순차적 입력 데이터에 멀티 헤드 어텐션을 적용하여 입력 데이터 내의 상관관계로부터 의미 있는 정보를 추출하는 트랜스포머 기술은, 최근 들어 대용량의 데이터를 사용하는 다양한 분야에서 성공적인 결과를 계속해서 거두고 있다(Vaswani et al., 2017). 음성인식 분야에서도 트랜스포머를 이용한 종단형 음성인식 기술은, 종래의 은닉 마코프 모델

* This work was supported by the National Research Foundation of Korea (NRF) and the Commercialization Promotion Agency for R&D Outcomes (COMPA) grant funded by the Korea government (MSIT) (RS-2023-00237117).
** pkyoung@etri.re.kr, Corresponding author
Received 16 May 2024; Revised 23 July 2024; Accepted 24 July 2024

과 딥러닝을 결합한 하이브리드 음성인식 기술의 한계를 극복하고 꾸준히 새로운 최고 성능을 갱신하고 있다(Dong et al., 2018).

본 논문에서는 트랜스포머를 베이스라인으로 하여, 이를 개선하여 음성인식에 특히 강점이 있다고 알려진 컨포머와 개선된 브랜치포머 모델을 한국어 음성인식에 적용하고, 각 모델의 성능을 비교한다(Gulati et al., 2020; Kim et al., 2023).

먼저 2절에서는 트랜스포머와 컨포머, E-브랜치포머를 간단히 설명하고 실험에 사용한 인코더-디코더 구조에 대해서 설명한다. 이후 3절에서는 실험에 사용한 한국어 훈련셋 및 평가셋에 대해서 설명한다. 4절에서는 앞서 소개한 모델과 데이터셋을 이용하여 각 모델을 훈련하고, 평가 및 분석한 결과를 소개하며, 5절에서 실험의 결과를 정리하고 결론을 맺는다.

2. 종단형 음성인식 모델

본 논문에서는 어텐션 기반의 인코더-디코더(AED) 구조의 종단형 음성인식 모델을 다룬다. 인코더로는 트랜스포머, 컨포머와 E-브랜치포머를 사용하였으며, 디코더는 모두 동일하게 트랜스포머를 사용하였다. AED 구조의 종단형 모델에서 인코더는 입력 시퀀스에 대하여 셀프 어텐션 기법을 적용하여 입력된 토큰 간의 관계를 반영함으로써, 중요한 정보를 추출하는 것으로 알려져있다. 또한 트랜스포머가 전체적인 입력 시퀀스의 글로벌한 정보를 잘 추출하는 반면 컨포머는 로컬한 정보 또한 잘 추출하도록 설계되었으며, 이러한 점이 이 두 모델이 음성인식에서 특히 잘 동작하는 장점으로 알려져있다. 각 모델에 대해서 간단히 설명하면 다음과 같다.

트랜스포머 인코더는 입력에 대해 positional encoding을 적용하여 위치 정보를 반영해준다. 이는 트랜스포머 인코더의 첫 번째 레이어로 전달된다. 트랜스포머 인코더는 여러 개의 인코더 레이어(또는 block)으로 구성되어 있는데, 각 레이어에서는 멀티 헤드 어텐션 연산을 수행하고 피드포워드 네트워크(FFN)을 거친 뒤 연산 결과를 다음 레이어에 전달해준다. 마지막 레이어의 결과는 트랜스포머 디코더의 입력으로 전달된다.

컨포머 인코더는 relative positional encoding을 적용하여 위치 정보를 반영해준다(Shaw et al., 2018). 이는 트랜스포머처럼, 인코더의 첫 번째 레이어로 전달된다. 컨포머 인코더 또한 여러 개의 인코더 레이어로 구성되어 있으며, 연산 결과가 다음 레이어에 전달되는 방식은 같다. 그러나 각 레이어에는 트랜스포머와 다르게 로컬한 정보를 잘 포착하는 컨볼루션 모듈이 추가되었다. 멀티 헤드 어텐션의 연산 결과가 컨볼루션 모듈에 전달되게 설계되었다. 또한, 멀티 헤드 어텐션 연산 이전에 FFN을 적용하고 컨볼루션 모듈 연산 이후에 FFN을 추가하는 macaron style FFN을 도입하였다. 그 결과, 컨포머는 음성인식에서 트랜스포머를 능가하는 성능을 보여주었다.

E-브랜치포머 인코더도 마찬가지로 여러 개의 인코더 레이어를 갖는다. 그러나 각 레이어의 구조는 트랜스포머, 컨포머와 다르다. E-브랜치포머는 글로벌한 정보를 추출하는 브랜치와 로컬한 정보를 추출하는 브랜치로 구성되어 있다. 전자는 멀티 헤드 어텐션 모듈로 구성되어 있고, 후자는 선형 레이어와 컨볼루션 모듈로 구성되어 있다. 각 브랜치에서 연산된 결과는 concatenate 연산으로 병합된다. 이는 브랜치포머 모델의 구조와 유사하다(Peng et al., 2022). E-브랜치포머는 브랜치포머 모델의 병합 모듈을 개선한 것으로, concatenate 연산 이후 컨볼루션 모듈을 추가하는 방식을 채택했다. 또한, 브랜치 분기 이전과 병합 모듈 이후에 FFN을 추가하여 컨포머 모델처럼 macaron style FFN을 도입하였고, 컨포머를 능가하는 성능을 보여주었다(Peng et al., 2023).

3. 데이터셋

모델의 훈련을 위해서 Oh et al.(2023)과 같이 AI Hub 데이터 7,500시간이 사용되었다. 훈련에 사용한 데이터의 전사데이터에는 한글과 영어, 숫자 뿐만 아니라 .,?!와 같은 문장부호, 그리고 +-~와 같은 기호도 포함되어 있으며, 이러한 기호들은 별도의 전처리 없이 그대로 훈련에 사용되었다.

표 1. 데이터셋별 발화 평균 길이(sec.)
Table 1. Average utterance length by dataset (sec.)

Dataset	Average length
fleurs-ko	12.58 (±3.70)
kmsav	6.59 (±5.16)
evalclean	3.17 (±2.86)
evalother	4.56 (±3.56)

표 2. 실험에 사용한 트랜스포머(T), 컨포머(C) 및 E-브랜치포머(E)의 모델 파라미터
Table 2. Model parameters of the Transformer (T), Conformer (C), and E-Branchformer (E) used in the experiment

Parameter	Model	Value
Num of blocks	T, C, E	12
Attention head		8
Attention dim		512
Input layer type		Conv2d
FFN dim		2048
Positional encoding type	T	Absolute
	C, E	Relative
Self-attention type	T	Global
	C, E	Relative

표 3. 실험에 사용한 트랜스포머(T), 컨포머(C) 및 E-브랜치포머(E)의 모델 크기
Table 3. Model sizes of the Transformer (T), Conformer (C), and E-Branchformer (E) used in the experiment

Model	Model size (MB)	Number of parameters (M)
Transformer (T)	292.91	73.23
Conformer (C)	445.13	111.28
E-Branchformer (E)	448.4	112.10

각 모델의 특성을 비교 분석하기 위하여 여러 공개 데이터를 평가에 사용하였다. Fleurs 데이터셋은 구글에서 공개한 데이터

셋으로 다국어 음성인식 평가에 널리 사용되며, 본 연구에서는 한국어 음성 부분만을 사용(fleurs-ko)하였다(Conneau et al., 2023). kmsav 데이터셋은 한국어 Youtube 데이터에서 수집된 멀티모달 데이터셋으로 발화별로 분할된 오디오 평가셋을 사용하였다(Park et al., 2024). evalclean 및 evalother는 KsponSpeech 데이터셋 중 평가데이터이며 KsponSpeech는 1,000시간 분량의 한국어 비정형 자유대화 음성데이터로, 훈련 데이터에도 사용된 데이터셋이다(Bang et al., 2020). 각 데이터셋의 발화별 평균 길이는 표 1에 표시하였다. 발화길이는 종단형 음성인식기의 성능과 속도에 영향을 크게 미치는 요소이므로, 다양한 길이의 발성을 대표하는 평가셋으로 선정하였다. 특히 kmsav 데이터셋은 평균 길이는 fleurs-ko 데이터셋보다 짧지만, 길이의 분산은 크며 30초 이상의 발화도 6발화가 포함되어 있으며, 가장 긴 발화는 35.84초에 달한다.

평가셋에 포함된 정답전사에는 한글, 영어, 숫자와 여러 기호들이 포함되어있으며, 문장 부호의 경우에는 평가셋마다 포함여부가 상이하였다. 따라서 본 실험에서는 인식 성능을 평가할 때 정답과 인식결과에서 모두 .,?!의 네 개의 기호만 제거하고 평가를 수행하였다. 또한 상황에 따라 동일한 발성이 영어나 숫자로 인식되거나, 한글의 형태로 표시되는 경우도 있다. 이러한 경우는 학습데이터의 유형에 따라 결정되는 것으로 가독성이 높은 인식결과를 얻기 위해서는 표기 전사의 형태로 출력되는 것이 바람직하다(Choi et al., 2024). 본 실험에서는 이와 같이 인식되는 형태에 따른 오류는 별도의 전처리 없이 그대로 포함되어 측정되었다.

4. 실험 및 분석

4.1. 인코더 모델에 따른 인식 성능

앞서 설명한 약 7,500시간의 한국어 데이터셋을 이용하여 각 모델을 훈련하고 평가셋을 이용하여 평가하였다. 동일한 환경에서 실험하기 위해 ESPnet에 공개된 recipe를 이용했다(Watanabe et al., 2018). 각 인코더 모델의 구체적인 설정값은 표 2와 같으며, 크기는 표 3에 표시하였다.

추가로 컨포머의 경우에는 macaron style FFN을 사용하였고, 31차원의 CNN 커널을 사용하였으며, E-브랜치포머의 경우 3072 차원의 MLP와 31차원의 CNN 커널, 31차원의 컨벌루션 커널을 이용하여 글로벌과 로컬 정보를 머지하였다.

표 4. 트랜스포머(T), 컨포머(C), E-브랜치포머(E) 세 가지 인코더 모델별 각 데이터셋에 대한 음절 오류율(%)
Table 4. Character error rate (%) for each dataset by encoder model: Transformer (T), Conformer (C), and E-Branchformer (E)

	fleurs-ko	kmsav	evalclean	evalother
T	4.54 (±0.03)	9.38 (±0.10)	6.83 (±0.03)	7.30 (±0.02)
C	4.30 (±0.04)	10.21 (±0.35)	6.54 (±0.06)	6.95 (±0.05)
E	4.11 (±0.11)	8.62 (±0.07)	6.17 (±0.00)	6.73 (±0.06)

The mean and standard deviation are shown for two models of the same structure trained with different initial values.

표 4는 트랜스포머, 컨포머 및 E-브랜치포머를 인코더로 사용했을 경우 각 모델별 인식결과이다. 음절단위의 토큰을 사용하였으며, 각 모델은 동일한 설정에 대하여 서로 다른 2개의 시드값으로 초기화하여 훈련하였으며 실험 결과는 2개의 모델에 대한 각각의 결과를 평균한 값을 표시하였다. 일반적으로 3개 이상의 시드로 학습하여 평균과 분산을 구하나, 실험의 종류가 많아서 2개의 시드에 대해서만 실험을 수행하였다. 각 시드에 따른 모델의 성능의 편차가 크지 않으므로 실험 결과는 신뢰 가능한 수준으로 볼 수 있다. 각 모델의 훈련은 50 epoch를 고정하여 훈련한 후 마지막 5개의 모델을 평균하여 사용하였다.

표 4에서 볼 수 있듯이 인식 성능은 E-브랜치포머가 가장 우수하며, 컨포머가 트랜스포머보다 우수한 것을 알 수 있으며, 초기값만을 다르게하여 훈련한 동일 구조의 모델 간의 성능 편차는 크지 않음을 알 수 있다. 표의 실험 결과 중 kmsav 데이터셋에 대한 컨포머 성능이 트랜스포머보다 떨어지는 것으로 나타나며, 이와 관련하여서는 4.4절에서 추가로 논의한다.

4.2. 인코더 모델에 따른 인식 속도

E-브랜치포머가 인식 성능은 가장 우수하나, 구조적으로 기존의 셀프 어텐션 구조와 로컬 브랜치를 병렬적으로 유지함으로써 계산량이 늘어날 것으로 예상할 수 있다. 이러한 계산량은 GPU 카드를 사용하거나, 고성능의 다코어 CPU를 사용하는 경우 병렬적으로 계산이 되어 전체적인 인식 속도에 영향을 주지 않을 수도 있으나, 저사양의 컴퓨팅 환경에서는 실사용에 제약을 줄 수도 있다. 이러한 점을 평가하기 위하여 저사양의 CPU를 가진 라즈베리파이 단말에서 각 모델의 인식 속도를 평가하였다.

표 5. 트랜스포머(T), 컨포머(C) 및 E-브랜치포머(E)의 세가지 인코더 모델에 따른 인식 속도(xRT)
Table 5. Recognition Speed (xRT) by encoder model: Transformer (T), Conformer (C), and E-Branchformer (E)

Model	Encoder only, beam=1		Encoder&decoder, beam=3	
	10s	20s	10s	20s
T	0.71 (±0.01)	1.13 (±0.01)	2.50 (±0.05)	3.79 (±0.05)
C	0.81 (±0.00)	1.23 (±0.01)	2.57 (±0.02)	3.85 (±0.02)
E	0.86 (±0.01)	1.27 (±0.03)	2.61 (±0.01)	3.92 (±0.01)

The recognition speed was measured using 10 randomly selected utterances from the KMSAV dataset, each approximately 10 and 20 seconds in length. The first two columns show the recognition speed with the CTC weight set to 1, using only the encoder and minimizing the computational load by setting the beam size to 1. The last two columns show the recognition speed in typical recognition scenarios.

실험에 사용한 단말은 라즈베리파이 5 단말로, 2.4 GHz의 클럭 스피드를 가진 4-core 64-bit Arm Cortex-A76 CPU가 장착되어 있다. 인식속도 평가를 위한 실험에서는 kmsav 테스트셋 중 약 10초 길이의 발화와, 약 20초 길이의 발화를 각각 무작위로 10개씩 선정하여 인식에 걸린 시간을 계산하고 이를 총 발화의 길이로 나누어 실시간 배율(real-time factor, xRT)을 계산하여 표 5에 나타내었다. xRT 값은 각 모델별로 두 개의 모델에 대하여 세 번씩 수행하여, 총 6번의 시행에 대한 평균값과 표준편차를

제시하였다. 속도를 평가하기 위한 디바이스로 저사양의 단말을 선택한 것은 각 모델 간의 속도차이를 극대화하여 보여주기 위한 것으로 높은 사양의 서버급 컴퓨터에서도 유사한 정도의 차이를 보여주었다.

표 5에서 encoder only, beam=1의 컬럼은 인코더 모델의 고유한 속도를 보다 정확히 측정하기 위하여 CTC 가중치를 1로 둠으로써 디코더를 사용하지 않고, 빔탐색에서 빔크기를 1로 두어 계산량을 최소로 둔 경우이며, encoder&decoder, beam=3의 컬럼은 일반적인 인식 환경과 동일하게 CTC 가중치를 0.3으로 두고 빔크기는 3으로 둔 경우이다. 앞의 컬럼에서 결과를 보듯이 인코더 구조를 E-브랜치포머로 변경함으로써, 연산속도가 느려지기는 하지만, 실제 인식환경에서는 디코더의 연산량이 대다수를 차지하며, 모델의 차이에 따른 인식 속도의 차이는 크지 않음을 확인하였다. 다만, 이러한 결과는 인식이 이루어지는 단말의 환경에 따라 큰 차이가 있음을 주의하여야 한다.

4.3. 한국어 토큰화에 따른 인식 성능

일반적으로 텍스트를 다루는 트랜스포머 모델의 경우 토큰화를 위하여 BPE(byte pair encoding)을 사용한다. 토큰으로 BPE 단위를 사용하는 경우, 단어를 사용하는 경우에 비해서 적은 유닛의 개수로 out-of-vocabulary 없이 모든 텍스트를 표현할 수 있다는 장점이 있으며, 모델의 크기나 활용 목적에 따라 토큰의 개수를 임의로 설정할 수 있는 장점 또한 가지고 있다.

영어의 경우 초기 종단형 모델의 경우 토큰 단위로 글자(character)를 사용하기도 하였으나, 영어의 경우 글자의 수가 100개 이내로 매우 작은 문제가 있다. 이와 비교하여 한국어의 경우 글자의 개수가 표기음절을 기준으로 3,000개 정도로 일반적으로 사용하는 BPE 크기와 유사하므로 음절을 BPE 단위로 사용하는 것도 가능한 선택이라고 할 수 있다.

표 6. 각 데이터셋에 대한 BPE 토큰 개수별 음절 오류율(%)
Table 6. Character error rate (%) by number of BPE tokens for each dataset

Num. BPE	Fleurs-ko	Kmsav	Evalclean	Evalother
Char	4.30 (±0.04)	10.21 (±0.35)	6.54 (±0.06)	6.95 (±0.05)
5 k	4.51 (±0.13)	10.20 (±0.13)	6.63 (±0.08)	7.01 (±0.09)
10 k	4.29 (±0.05)	10.08 (±0.11)	6.52 (±0.02)	7.05 (±0.08)
15 k	4.29 (±0.10)	9.63 (±0.21)	6.48 (±0.07)	6.93 (±0.03)
20 k	4.29 (±0.14)	9.70 (±0.38)	6.45 (±0.04)	7.04 (±0.06)

The mean and standard deviation of the recognition results are shown for two Conformer models of the same structure trained with different initial values.

표 7. BPE 토큰 개수 별 인식 속도(xRT)
Table 7. Recognition Speed (xRT) by number of BPE tokens

Num. BPE	10s	20s
Char	2.57 (±0.02)	3.85 (±0.02)
5 k	1.46 (±0.02)	2.19 (±0.02)
10 k	1.28 (±0.01)	1.89 (±0.02)
15 k	1.22 (±0.01)	1.78 (±0.01)
20 k	1.21 (±0.02)	1.70 (±0.01)

The first column shows the average xRT for 10 utterances each approximately 10 seconds long, and the second column shows the average xRT for 10 utterances each approximately 20 seconds long.

표 6에서는 토큰의 개수에 따른 인식성능을 비교하였다. 현재도 가장 많이 사용되는 음성인식 모델은 컨포머이므로, 컨포머 모델에 대해서 실험을 수행했다.

음절을 토큰으로 사용한 경우와 BPE 유닛의 개수를 5 k, 10 k, 15 k, 20 k로 달리하여 사용한 경우 음성인식 성능에 있어 유의미한 차이는 발생하지 않았으나, 초기값을 다르게하여 훈련한 두 모델의 성능편차는 유닛의 개수가 많아짐에 따라 커짐을 확인할 수 있었다.

표 7은 토큰의 개수에 따른 인식 속도를 비교한 표이다. BPE 유닛의 개수가 커질수록 하나의 토큰의 길이가 길어지고, 결과적으로 한 문장을 표현하는데 필요한 토큰의 개수가 줄어든다. 따라서 AED 구조에서 디코더의 연산량이 그만큼 감소하므로 추론 속도가 빨라질 것을 예상할 수 있으며, 표에서 이를 확인할 수 있었다. 단 BPE 토큰 개수의 늘어남에 따라, 출력층에서의 softmax 계산량도 증가하므로 이에 따라, BPE 토큰 개수 증가에 따른 속도 개선 효과는 크지 않았다.

4.4. 긴 발화에 대한 컨포머 모델 성능

앞서 표 4의 결과 중 컨포머의 성능은 대부분의 데이터셋에서 트랜스포머보다 우수한 성능을 보이나, kmsav 데이터셋에 대해서는 트랜스포머보다 열등한 성능을 보인다. 이에 대한 원인을 분석하기 위하여 kmsav 데이터셋에 대하여 발화길이 별로 구분하여 인식성능을 표 8에 나타내었다. 일반적으로 트랜스포머 모델은 길이가 길어질수록 인식속도가 느려진다고 알려져 있으며 이는 셀프 어텐션의 계산량이 길이의 제곱에 비례하여 늘어나기 때문이다(Child et al., 2019). 또한 성능 또한 길이가 길어짐에 따라 저하되는 것이 일반적이다. 컨포머의 경우 길이에 따른 성능 차이가 특히 두드러진다. 예를 들어 25초 이상의 길이의 발화에 대해서 트랜스포머 및 E-브랜치포머는 오류율이 6.65%–16.05%의 범위에 있으나, 컨포머에 대해서는 32.27%–70.12%까지 저하되는 것을 확인할 수 있다.

표 8. kmsav 데이터셋 중 트랜스포머, 컨포머, E-브랜치포머에 대한 발화 길이별 인식성능(CER, %)

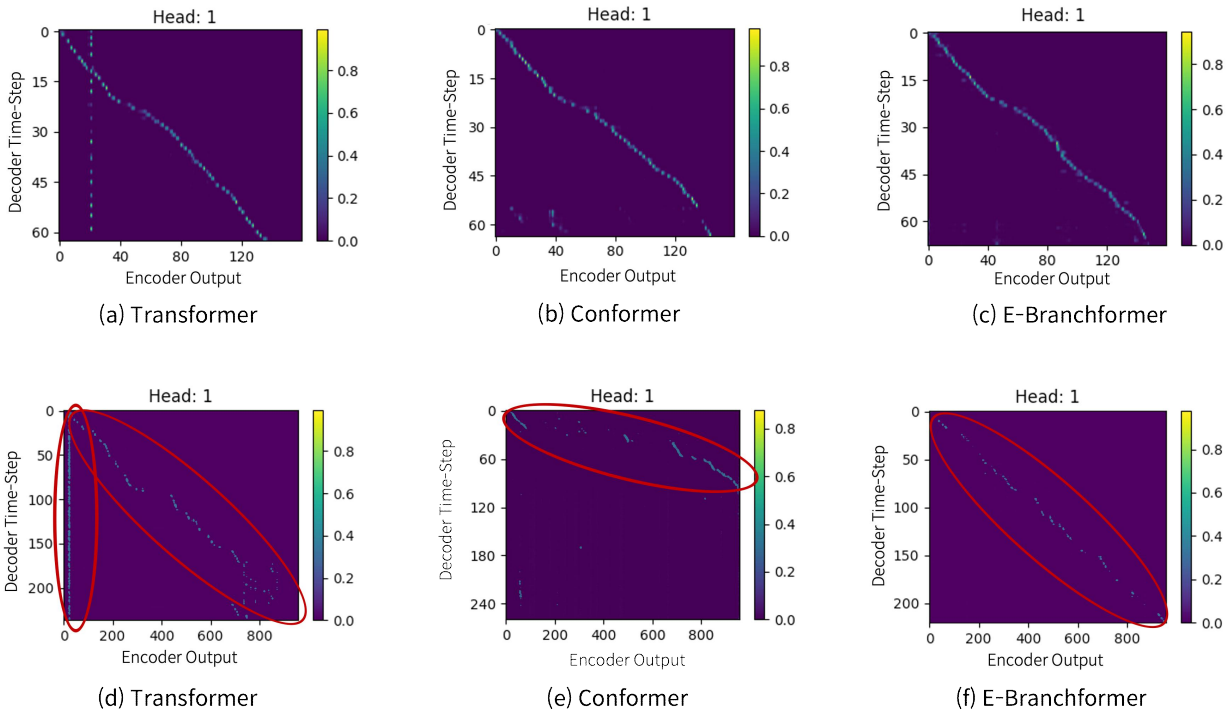
Table 8. Recognition performance (CER, %) by utterance length for Transformer, Conformer, and E-Branchformer on the KMSAV Dataset

발화길이(초)	Transformer	Conformer	E-Branchformer
0–5	17.77 (±0.15)	16.82 (±0.41)	15.94 (±0.15)
5–10	8.18 (±0.16)	8.05 (±0.08)	7.79 (±0.09)
10–15	6.86 (±0.02)	6.74 (±0.25)	6.36 (±0.09)
15–20	6.44 (±0.03)	6.10 (±0.06)	5.83 (±0.01)
20–25	7.79 (±0.13)	12.25 (±2.66)	7.03 (±0.08)
25–30	16.05 (±1.12)	32.27 (±14.37)	11.01 (±2.80)
30–35	10.84 (±0.56)	46.32 (±3.86)	11.03 (±0.00)
35–40	9.47 (±0.54)	70.12 (±6.56)	6.65 (±0.64)
All	9.38 (±0.10)	10.21 (±0.35)	8.62 (±0.07)

컨포머에 대한 긴 발화 인식 성능 저하 현상은 Pan et al. (2022)에서도 언급된 바 있다. 다만 원인에 대한 설명은 언급되지 않았다. 성능 저하의 원인을 확인하기 위하여 인코더-디코더 간의 크로스 어텐션을 분석해보았다. 그림 1은 각 모델별로 크로스 어텐션을 2차원 히트맵으로 표시한 것이다. 각 히트맵의 x축은 인코더의 입력 특징벡터 인덱스, y축은 디코더의 출력 토큰 인덱스를 나타내며, 맵의 밝은 부분이 해당 토큰을 출력하는데 기여한 입력 특징벡터의 인덱스를 표시한다. 첫 번째 줄의 세 개의 그림은 인식이 정상적으로 이루어진 짧은 발화에 대한 히트맵으로 모든 입력 특징벡터가 인식결과를 출력하는데 사용되었음을 확인할 수 있다. 이에 반하여 아래 줄의 세 개의 히트맵은 길이가 긴 발화에 대한 크로스 어텐션 스코어를 나타낸 것으로 이 중 가운데의 것이 인식 오류가 발생한 컨포머에 대한 것이다. 인식 결과 중 앞부분을 출력하는데 대부분의 입력 특징벡터가 소모되었고, 이후 의미없는 결과가 출력되었다.

트랜스포머 및 E-브랜치포머와 비교하여 컨포머의 경우에만 이러한 현상이 발생하는 구체적인 기전에 대해서는 확인하지 못하였으나, 두 모델과 비교하여 컨포머의 경우에는 컨볼루션 연산을 통하여 부분적인 특징을 강조함에 따라 현재의 결과를 출력하는데 최적인 입력특징벡터를 전체 프레임에서 더 적극적으로 찾아감에 따라 입력과 출력 간의 정렬이 제대로 이루어지지 못한 것으로 추정된다. 이러한 문제를 해결하기 위해서는 컨포머의 구조에서 시간적으로 가까운 부분만을 이용하여 결과를 출력하도록 제한하는 구조가 필요할 것으로 생각되며 이에 관한 연구가 진행 중에 있다.

컨포머에 대해서 발생하는 이러한 현상을 막기 위한 보다 간단한 해결 방법으로는 발화를 20초 이내의 비교적 짧은 길이로 분할하여 인식하는 것이 필요하다. 예를 들어 Bain et al. (2023)에서는 음성 검출기를 활용하여 긴 발화 중 휴지부를 검출하여 발화를 짧은 단위로 나눈 후 나누어진 발화들을 배치 인식하여 인식 속도를 크게 증가시키고 있다. 컨포머에 대한 이러한 방법



Figures (a), (b), and (c) illustrate cross-attention scores for short utterances, while figures (d), (e), and (f) represent those for long utterances. The short utterances are randomly selected from the evaluation set of KsponSpeech, and the long utterances are created by combining two randomly selected utterances from the same dataset, resulting in approximately 35 seconds of speech. All figures (a) to (f) visualize the cross-attention scores from the second head (index 1) of the last decoder layer (6th layer). For long utterances, visibility is reduced compared to short utterances, so high-score regions are marked separately.

그림 1. 트랜스포머, 컨포머 및 E-브랜치포머에 대한 짧은 발화 및 긴 발화에 대한 교차 어텐션 스코어

Figure 1. Cross-attention scores for short and long utterances for Transformer, Conformer, and E-Branchformer

은 속도 뿐만 아니라, 인식 성능 또한 개선할 수 있다.

5. 결론

본 논문에서는 한국어 종단형 음성인식 모델의 종류에 따른 음성인식 성능과 속도를 비교하였다. 언어처리, 영상 등 다양한 분야와 마찬가지로 음성인식 분야에서도 어텐션 기반의 인코더-디코더 모델이 널리 활용되고 있으며, 특히 음성인식 분야에서는 컨포머 및 E-브랜치포머 등 개선된 AED 모델이 제안되었다. 본 논문에서는 한국어 음성인식 도메인에서 각 모델의 성능과 속도를 실험적으로 측정하여 비교하였다.

또한 인식 단위로 음절을 사용하는 경우와 서브워드를 사용하는 경우를 비교하였으며 서브워드의 경우 Byte Pair Encoding의 토큰 수를 변화시키면서 인식 성능과 속도를 비교하였다.

실험결과 한국어에 있어서도 E-브랜치포머가 우수한 성능을 보였으며, 현재 널리 활용되고 있는 컨포머도 트랜스포머보다는 우수한 성능을 보이나, 길이가 긴 발화에 대해서는 특히 인식 성능이 저하됨을 확인하였다. 성능 저하의 원인으로는 인코더와 디코더의 크로스 어텐션 계산부에서 입력과 출력 간의 정렬 과정에서의 오차가 발생함도 확인하였다.

향후 연구에서는 이러한 실험 결과를 바탕으로 대규모 한국어 음성인식 모델의 학습에 활용하며, 컨포머의 인식오류를 개선하는 연구를 진행하고자 한다.

References

Bain, M., Huh, J., Han, T., & Zisserman, A. (2023, August). WhisperX: Time-accurate speech transcription of long-form audio. *Proceedings of the Interspeech 2023* (pp. 4489-4493). Dublin, Ireland.

Bang, J. U., Yun, S., Kim, S. H., Choi, M. Y., Lee, M. K., Kim, Y. J., Kim, D. H., ... Kim, S. H. (2020). KsponSpeech: Korean spontaneous speech corpus for automatic speech recognition. *Applied Sciences*, *10*(19), 6936.

Child, R., Gray, S., Radford, A., & Sutskever, I. (2019). Generating long sequences with sparse Transformers. *arXiv*. https://doi.org/10.48550/arXiv.1904.10509.

Choi, H., Choi, M., Kim, S., Lim, Y., Lee, M., Yun, S., Kim, D., ... Kim, S. H. (2024). Spoken-to-written text conversion for enhancement of Korean-English readability and machine translation. *ETRI Journal*, *46*(1), 127-136.

Conneau, A., Ma, M., Khanuja, S., Zhang, Y., Axelrod, V., Dalmia, S., Riesa, J., ... Bapna, A. (2023, January). Fleurs: Few-shot learning evaluation of universal representations of speech. *Proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT)* (pp. 798-805). Doha, Qatar.

Dong, L., Xu, S., & Xu, B. (2018, April). Speech-Transformer: A no-recurrence sequence-to-sequence model for speech recognition. *Proceedings of the 2018 IEEE International Conference on Acoustics*, *Speech and Signal Processing (ICASSP)* (pp. 5884-5888). Calgary, AB.

Gulati, A., Qin, J., Chiu, C. C., Parmar, N., Zhang, Y., Yu, J., Han, W., ... Pang, R. (2020, October). Conformer: Convolution-augmented Transformer for speech recognition. *Proceedings of Interspeech 2020* (pp. 5036-5040). Shanghai, China.

Kim, K., Wu, F., Peng, Y., Pan, J., Sridhar, P., Han, K. J., & Watanabe, S. (2023, January). E-Branchformer: Branchformer with enhanced merging for speech recognition. *Proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT)* (pp. 84-91). Doha, Qatar.

Oh, C., Kim, C., & Park, K. (2023). Building robust Korean speech recognition model by fine-tuning large pretrained model. *Phonetics and Speech Sciences*, *15*(3), 75-82.

Peng, Y., Dalmia, S., Lane, I., & Watanabe, S. (2022, June). Branchformer: Parallel mlp-attention architectures to capture local and global context for speech recognition and understanding. *Proceedings of the International Conference on Machine Learning* (pp. 17627-17643). Baltimore, MD.

Peng, Y., Kim, K., Wu, F., Yan, B., Arora, S., Chen, W., Tang, J., ... Watanabe, S. (2023, August). A comparative study on E-Branchformer vs Conformer in speech recognition, translation, and understanding tasks. *Proceedings of Interspeech 2023* (pp. 2208-2212). Dublin, Ireland.

Pan, J., Lei, T., Kim, K., Han, K. J., & Watanabe, S. (2022, May). SRU++: Pioneering fast recurrence with attention for speech recognition. *Proceedings of the 2022 IEEE International Conference on Acoustics*, *Speech and Signal Processing (ICASSP)* (pp. 7872- 7876). Singapore, Singapore.

Park, K., Oh, C., & Dong, S. (2024). KMSAV: Korean multi-speaker spontaneous audiovisual dataset. *ETRI Journal*, *46*(1), 71-81.

Shaw, P., Uszkoreit, J., & Vaswani, A. (2018, June). Self-attention with relative position representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *Volume 2 (Short Papers)* (pp. 464-468). New Orleans, Louisiana.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł, ... Polosukhin, I. (2017, Deccember). Attention is all you need. *Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017)*. Long Beach, CA.

Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Enrique Yalta Soplin, N., ... Ochiai, T. (2018, September). ESPnet: End-to-end speech processing toolkit. *Proceedings of the Interspeech 2018* (pp. 2207-2211). Hyderabad, India.

• **오창한 (Changhan Oh)**
한국전자통신연구원 복합지능연구실 UST 학생연구원
대전광역시 유성구 가정로 218
Tel: 042-860-6114
이메일: ochh508@etri.re.kr
관심분야: 인공지능, 음성인식, 복합지능

• **김민서 (Minseo Kim)**
한국외국어대학교 ELLT학과
서울 동대문구 이문로 107
Tel: 042-860-6114
이메일: er1123090@gmail.com
관심분야: 인공지능, 음성인식, 복합지능

• **박기영 (Kiyoung Park)** 교신저자
한국전자통신연구원 복합지능연구실 책임연구원
대전광역시 유성구 가정로 218
Tel: 042-860-1228
이메일: pkyoung@etri.re.kr
관심분야: 인공지능, 음성인식, 복합지능

• **송화전 (Hwajeon Song)**
한국전자통신연구원 복합지능연구실 책임연구원
대전광역시 유성구 가정로 218
Tel: 042-860-5836
이메일: songhj@etri.re.kr
관심분야: 인공지능, 음성인식, 복합지능

트랜스포머 기반 모델의 한국어 음성인식 성능 비교 연구*

오 창 한[1] · 김 민 서[2] · 박 기 영[1] · 송 화 전[1]

[1]한국전자통신연구원 복합지능연구실, [2]한국외국어대학교 ELLT학과

국문초록

트랜스포머 모델은 텍스트, 영상 등 순차적 입력 데이터에서 의미 있는 정보를 추출하는 데 뛰어난 성과를 보여주었으며, 음성인식 분야에서도 종단형 모델로서 주목받고 있다. 본 연구에서는 트랜스포머 음성인식 모델과 이를 개선한 컨포머, E-브랜치포머 모델을 한국어 음성인식에 적용하여 성능을 비교하였다. AIHub에 공개된 한국어 음성 데이터를 활용하여 약 7,500시간의 훈련셋을 마련하고, ESPnet 툴킷을 활용하여 트랜스포머, 컨포머, E-브랜치포머 모델을 훈련하고 성능을 평가하였다. 또한, 인식 단위로 음절과 서브워드를 사용하는 경우를 비교하고, Byte Pair Encoding의 토큰 수 변화에 따른 성능 차이를 분석하였다. 실험 결과, E-브랜치포머가 한국어 음성인식에서 가장 우수한 성능을 보였으며, 컨포머는 트랜스포머보다 우수하였으나 긴 발화에 대해서는 성능 저하가 확인되었다. 이러한 성능 저하의 원인으로 인코더-디코더의 크로스 어텐션 정렬 과정에 오차가 발생함을 확인하였다. 또한, 서브워드 인식 단위를 사용하면서 토큰 수를 조정할 때의 성능 변화에 대한 분석을 통해 최적의 설정을 찾고자 하였다. 본 연구는 모델의 정확도와 처리 속도를 종합적으로 평가하였으며, 이를 통해 한국어 음성인식의 효율성을 극대화할 수 있는 방법을 모색하였다. 대규모 한국어 음성인식 모델의 학습과 컨포머의 인식 오류 개선 연구에 기여할 수 있을 것으로 기대된다. 또한, 향후 연구 방향으로는 다양한 한국어 음성 데이터셋을 활용한 추가 실험과 더불어, 컨포머의 구조적 개선을 통한 인식 성능 향상을 목표로 한다.

핵심어: 딥러닝, 머신러닝, 음성인식, 음성공학

참고문헌

오창한, 김청빈, 박기영 (2023). 대형 사전훈련 모델의 파인튜닝을 통한 강건한 한국어 음성인식 모델 구축. *말소리와 음성과학*, *15*(3), 75-82.

* 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단과 과학기술사업화진흥원의 지원을 받아 수행된 연구임(RS-2023-00237117).