한국음향학회지 제40권 제5호 pp. 530~536 (2021)
The Journal of the Acoustical Society of Korea Vol.40, No.5 (2021)
https://doi.org/10.7776/ASK.2021.40.5.530

pISSN : 1225-4428
eISSN : 2287-3775

Lexicon transducer를 적용한 conformer 기반 한국어 end-to-end 음성인식

Conformer with lexicon transducer for Korean end-to-end speech recognition

손현수,[1] 박호성,[1] 김규진,[1] 조은수,[1] 김지환[1†]

(Hyunsoo Son,[1] Hosung Park,[1] Gyujin Kim,[1] Eunsoo Cho,[1] and Ji-Hwan Kim[1†])

[1]서강대학교

(Received August 10, 2021; accepted September 7, 2021)

초 록: 최근 들어 딥러닝의 발달로 인해 Hidden Markov Model(HMM)을 사용하지 않고 음성 신호와 단어를 직접 매핑하여 학습하는 end-to-end 음성인식 방법이 각광을 받고 있으며 그 중에서도 conformer가 가장 좋은 성능을 보이고 있다. 하지만 end-to-end 음성인식 방법은 현재 시점에서 어떤 자소 또는 단어가 나타날지에 대한 확률에 대해서만 초점을 두고 있다. 그 이후의 디코딩 과정은 현재 시점에서 가장 높은 확률을 가지는 자소를 출력하거나 빔 탐색을 사용하며 이러한 방식은 모델이 출력하는 확률 분포에 따라 최종 결과에 큰 영향을 받게 된다. 또한 end-to-end 음성인식 방식은 전통적인 음성인식 방법과 비교 했을 때 구조적인 문제로 인해 외부 발음열 정보와 언어 모델의 정보를 사용하지 못한다. 따라서 학습 자료에 없는 발음열 변환 규칙에 대한 대응이 쉽지 않다. 따라서 본 논문에서는 발음열 정보를 담고 있는 Lexicon transducer(L transducer)를 이용한 conformer의 디코딩 방법을 제안한다. 한국어 데이터 셋 270 h에 대해 자소 기반 conformer의 빔 탐색 결과와 음소 기반 conformer에 L transducer를 적용한 결과를 비교 평가하였다. 학습자료에 등장하지 않는 단어가 포함된 테스트 셋에 대해 자소 기반 conformer는 3.8 %의 음절 오류율을 보였으며 음소 기반 conformer는 3.4 %의 음절 오류율을 보였다.

핵심용어: 음성인식, 트랜스포머, Weighted finite state transducer, End-to-end

ABSTRACT: Recently, due to the development of deep learning, end-to-end speech recognition, which directly maps graphemes to speech signals, shows good performance. Especially, among the end-to-end models, conformer shows the best performance. However end-to-end models only focuses on the probability of which grapheme will appear at the time. The decoding process uses a greedy search or beam search. This decoding method is easily affected by the final probability output by the model. In addition, the end-to-end models cannot use external pronunciation and language information due to structual problem. Therefore, in this paper conformer with lexicon transducer is proposed. We compare phoneme-based model with lexicon transducer and grapheme-based model with beam search. Test set is consist of words that do not appear in training data. The grapheme-based conformer with beam search shows 3.8 % of CER. The phoneme-based conformer with lexicon transducer shows 3.4 % of CER.

Keywords: Speech recognition, Transformer, Weighted finite state transducer, End-to-end

PACS numbers: 43.72.Ne, 43.72.Bs

†**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)

I. 서 론

음성인식은 음성 시퀀스를 단어의 시퀀스로 변경하는 과정이다. 하지만 음성 시퀀스의 길이와 단어 시퀀스의 길이는 서로 다르다. 따라서 딥러닝 학습을 위해 입력 음성에 대응되는 출력 시퀀스를 정렬할 필요가 있으며 이를 정렬 문제라 한다. 전통적인 음성인식 방법은 이를 Hidden Markov Model(HMM)을 이용하여 해결하였지만 HMM을 사용하기 위해서는 여러 가지 가정과 해당 언어데 대한 지식을 필요로 한다. 하지만 딥러닝의 발전에 따라 입력 신호를 직접 단어의 자소에 매핑하여 HMM을 통한 사전 정렬이 필요 없는 end-to-end 방법에 대한 연구가 활발히 이루어지며 최근에는 전통적인 DNN-HMM 방식의 음성인식 보다 좋은 성능을 보이고 있다. 그 중 시퀀스 모델링에 장점을 가진 Transformer[1]와 지역적인 특징을 잘 추출하는 콘볼루션(convolution)의 장점을 결합하여 입력 음성에 대해 우수한 특징 벡터를 뽑아내는 Convolution-Augmented Transformer(Conformer)가 가장 좋은 성능을 보이고 있다.[2]

하지만 conformer를 포함한 모든 end-to-end 모델들은 현재 시점에서 어떠한 자소가 나타날지에 대한 확률에 대해서만 초점을 두고 있으며 이후의 디코딩 과정은 가장 확률이 높은 라벨을 나열하거나 빔 탐색을 사용한다. 이러한 디코딩 방식은 모델이 출력하는 확률 분포에 따라서 최종 결과에 큰 영향을 미친다. 예를 들어, 정답이 'cat'인 발성에 대해 올바른 end-to-end 모델은 '__c__a__t__'라는 결과를 출력하게 된다. 하지만 하나의 프레임에 대해 올바른 확률 분포를 출력하지 못하였다면 '__c__a__e__t'를 출력하게 되고 이는 최종 결과가 'caet'가 된다. 이처럼 한 프레임의 오류지만 단어 전체가 오류가 되는 현상이 발생한다. 따라서 발생할 수 있는 시퀀스에 대한 경로를 그래프 형태로 미리 만들어 놓아 가능하지 않은 경로를 제한할 필요가 있다.

또한 end-to-end 방법은 전통적인 음성인식 방법과 비교해 볼 때 그 구조적인 문제로 인해 외부 발음열 정보를 사용하지 못한다. 특히 자소를 출력단으로 하는 end-to-end 모델의 경우 동일한 자소라도 앞뒤 문맥에 따라 다르게 발음되는 경우가 있다. 예를 들어, '학교'에 경우 이를 자소열로 나열하면 'ㅎㅏㄱㄱㅛ'이다. 이때 첫 번째 'ㄱ'과 두 번째 'ㄱ'은 동일한 소이지만 실제로는 다르게 발음된다. 이러한 경우 end-to-end 모델은 학습자료에 위와 같은 발음열 규칙이 포함된 문장이 없는 경우 올바른 정답을 출력하기 어렵다. 따라서 발음열 정보를 담고 있는 L transducer를 Weighted Finite State Transducer(WFST)[3]를 통해 미리 가능한 음소열에 대한 정보를 그래프화 하여 end-to-end 모델의 탐색 범위를 제한하여 앞서 언급한 문제를 해결하고자 한다. Tables 1과 2는 기존 end-to-end 음성인식에서 발생한 오류의 예를 보여준다. Table 1은 소수의 프레임에서 올바른 확률 분포를 출력하지 못한 경우의 예를 보여주고 Table 2는 올바른 발음열 변환이 되지 않은 경우를 보여준다.

Table 1. Error occurs in a small number of frames.

Hypothesis	정상회담이 매우 성공적혔다고 평가했습니다
Reference	정상회담이 매우 성공적이였다고 평가했습니다
Hypothesis	도쿄오쿄 영화제에서 특별 상영되었습니다
Reference	도쿄 영화제에서 특별 상영되었습니다
Hypothesis	누빙의 유적들이 전시되어 있습니다
Reference	누비아의 유적들이 전시되어 있습니다
Hypothesis	정상회담을 기념해 느내놓은 음반의 타이틀곡
Reference	정상회담을 기념해 내놓은 음반의 타이틀곡

Table 2. Pronunciation sequence is not poperly converted.

Hypothesis	페인팅족들이 엮어내는 콜로풀한 율동으로
Reference	페인팅족들이 엮어내는 컬러풀한 율동으로
Hypothesis	볼래 의미가 사라졌습니다
Reference	본래 의미가 사라졌습니다
Hypothesis	혜소는 실라말의 선승입니다
Reference	혜소는 신라말의 선승입니다
Hypothesis	유니상씨의 제자인 소프라노 윤인숙씨가
Reference	윤희상씨의 제자인 소프라노 윤인숙씨가

본 논문에서는 conformer의 출력단을 음소 단위로 변경하여 학습한 모델에 L transducer를 이용한 디코딩 방법을 제안한다. 제안하는 방법의 conformer 모델과 기존 자소 기반 conformer 모델은 출력단위만 다를 뿐 동일한 구조를 사용하였다. 한국어 데이터 270 h을 사용하여 학습하였으며 학습자료에 등장하지 않는 단어를 포함한 100개 문장에 대해 성능 비교

평가를 수행한다.

II. 관련 연구

오랜 기간 동안 음성인식에서는 HMM 기반 모델을 사용하였다. 일반적으로 HMM 기반 모델은 음향모델, 발음모델, 언어모델로 나눌 수 있으며, 각 부분은 서로 독립적이고 역할이 다르다. 음향모델은 입력 오디오와 HMM 상태 간의 매핑을 모델링 하며 언어모델은 단어 시퀀스를 문장으로 매핑한다. 이때 단어를 음소를 매핑하는 과정에서 발음 모델이 사용되며 발음 모델을 정의할 때는 타겟 언어에 대한 특별한 지식을 필요로 한다. 하지만 딥러닝의 발전에 따라 Deep Neural Network(DNN)이 음성인식에 적용되었고 HMM의 사후 확률을 계산하던 GMM의 역할이 DNN으로 교체되었으며 이를 DNN-HMM 기반 음성인식이라 한다. 하지만 DNN-HMM 기반 음성인식은 DNN의 한계로 인해 강제 정렬을 필요로 하고 각 요소들이 서로 독립적인 학습 가설을 가지고 있기 때문에 전체적인 최적화가 어렵다는 단점을 가지고 있다. 위와 같은 단점과 딥러닝 기술의 발전으로 점점 더 많은 연구가 end-to-end 방식으로 이루어졌다.

End-to-end 음성인식의 학습 철학은 DNN-HMM 기반 음성인식의 구성 요소인 음향 모델, 언어 모델, 발음 모델을 모두 하나의 네트워크로 학습하는데 있다. 즉, 입력 오디오 시퀀스를 자소나 단어 시퀀스로 직접 매핑하는 시스템이다. 대부분의 end-to-end 음성인식은 인코더와 디코더로 이루어져 있다. 인코더는 입력 시퀀스를 특징 벡터 시퀀스로 매핑하며 디코더는 특징 벡터 시퀀스를 출력 시퀀스로 매핑한다. End-to-end 모델의 가장 큰 특징은 디코더를 통해 추론 시 입력 시퀀스의 길이를 출력 시퀀스의 길이로 변경하며 학습 시에는 출력 시퀀스의 길이를 입력 시퀀스의 길이로 변경하여 딥러닝 모델의 학습이 가능하게 하여 정렬 문제를 해결하는데 있다. 이러한 end-to-end 모델은 크게 Connectionist Temporal Classification(CTC)[4]와 RNN Transducer(RNN-T)[5] 구조로 나눌 수 있다.

CTC의 가장 큰 장점 중 하나는 데이터 강제 정렬의 필요성을 제거하여 CNN 및 RNN과 같은 딥러닝 기술이 점점 더 중요한 역할을 할 수 있게 한 것이다. Alex Graves *et al*.[4]는 처음으로 CTC를 적용한 end-to-end 음성인식을 제안하였다.[4] 하지만 CTC는 출력 시퀀스의 라벨이 서로 독립적임을 가정하였다. 이러한 가정은 언어모델의 정보를 학습하지 못한다는 것을 의미하며 CTC로 학습된 모델은 음향모델의 역할 만을 수행할 수 있다는 것을 의미한다.

이러한 문제를 해결하기 위해 RNN-T가 제안되었다.[5] RNN-T는 전사 네트워크, 예측 네트워크, 결합 네트워크로 구성된다. CTC와의 차이점은 예측 네트워크로 인해 라벨의 시퀀스도 동시에 학습하여 언어모델의 역할을 하는데 있다.

이처럼 end-to-end 음성인식의 기존 연구는 다양한 모델들을 사용하여 현재 시점에서 출력 노드에 대한 가장 좋은 확률을 뽑아내는데 초점을 두고 있다. 이후의 최종 결과로의 디코딩은 현재 시점에서 가장 확률이 높은 출력열을 그대로 이어붙이는 greedy 알고리즘이나 몇 개의 후보를 남기는 beam 탐색 방법을 사용한다. 이러한 방식은 특정 라벨에 대한 확률이 한 프레임이라도 오류가 발생할 시 단어 전체의 오류로 이어지기 쉬운 구조를 가진다. 따라서 디코딩 네트워크를 미리 만들어 놓은 후 해당 경로에 대해서만 탐색하여 탐색 공간을 줄일 필요가 있다. 따라서 본 논문에서는 WFST를 이용하여 네트워크를 미리 구축하고 탐색 공간을 줄임과 동시에 프레임 오류로 인해 발생하는 인식결과의 오류를 해결할 수 있는 end-to-end의 디코딩 방식을 제안한다.

III. Lexicon Transducer를 이용한 디코딩 방법

FST는 상태 전이가 입력 심볼과 출력 심볼로 정의되는 유한 오토마타이다. 따라서 트랜스듀서를 통과하는 경로는 입력 심볼 시퀀스에서 출력 심볼 시퀀스로의 변환을 의미한다. 즉, FST는 가능한 모든 시퀀스를 표현한 그래프이다. WFST는 FST에 상태 전이의 가중치를 부여한 것을 말한다. 가중치는 입력 심볼과 출력 심볼로 변경하는 확률을 계산하기 위해 사용하여 시퀀스의 변경의 확률은 가중치의 곱으로 표현된다.

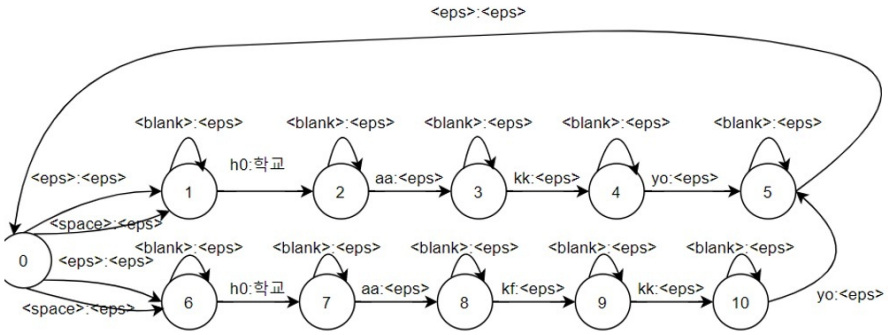

Fig. 1. Example of L transducer.

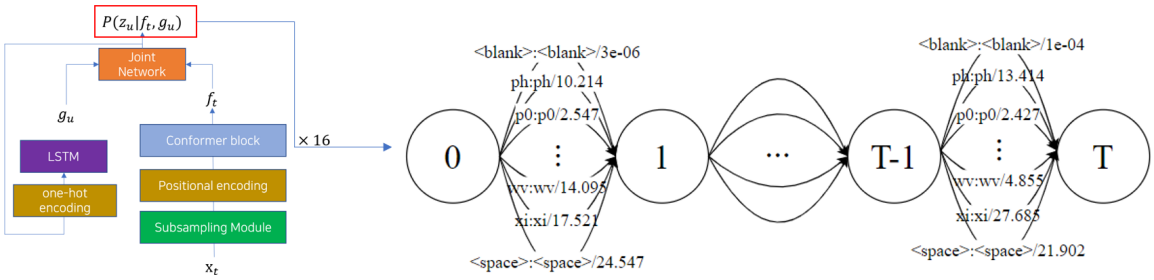

Fig. 2. (Color available online) Example of U transducer.

L transducer는 발음사전의 정보를 토대로 음소의 시퀀스를 단어로 변경하는 WFST이다. 발음사전은 정의된 단어에 대응되는 음소 시퀀스의 쌍으로 구성되며 이 정보를 WFST로 표현한 것이 L transducer이다. Fig. 1은 단어 '학교'에 대한 가능한 발음열 '하꾜'와 '학꾜'를 WFST로 만들었을 때 구조를 보여준다. 이때 end-to-end의 띄어쓰기와 <blank> 심볼을 처리하기 위해 FST의 상태마다 셀프 루프를 추가한다.

음소기반 end-to-end 모델은 입력 프레임 시퀀스를 음소 시퀀스로 매핑한다. 따라서 음소의 시퀀스를 단어의 시퀀스로 매핑할 필요가 있다. 이 디코딩 과정은 L transducer와 U transducer 2개의 WFST로 진행된다. L transducer는 음소 시퀀스를 단어 시퀀스로 변경하며 이 때, 음소 시퀀스에 매핑되는 단어는 Grapheme to Phoneme(G2P) 모델에 따라 결정된다. 본 논문에서 G2P는 규칙 기반의 한국어 G2P[6]를 사용하였다.

U transducer는 음소 기반 end-to-end 모델의 최종 출력 노드로부터 나온 각 음소에 대한 확률 값을 FST로 변경한 WFST이다. U transducer의 가중치는 end-to-end 모델의 최종 출력 확률 값에 −ln을 취한 값을 사용한다. Fig. 2는 U transducer의 예를 보여준다.

이후 U transducer와 L transducer를 composition 하여 디코딩 네트워크를 구축한다. 구축된 네트워크의 경로 중 가장 가중치가 작은 경로의 출력 값을 정답으로 하여 출력한다. 따라서 최종 결과는 다음과 같은 Eq. (1)으로 표현된다.

$$shortestpath(U \circ L). \tag{1}$$

이 때 shortestpath(X)는 FST의 가능한 경로 상 가장 가중치가 작은 경로를 말한다.

IV. 실 험

학습에 사용한 데이터는 한국전자통신연구원에서 구매한 약 276 h의 한국어 음성 코퍼스를 사용하였다. 해당 데이터는 조용한 사무실 환경에서 낭독체 문장을 발성하여 녹음한 음성 데이터로 약 2,000명의 화자가 발성하였다. 테스트 데이터는 학습자료에 등장하지 않는 단어를 포함한 100개의 문장을 직접 수집하여 사용하였다. 모든 음성파일을 16,000 샘플링, 샘플 당 2 바이트, 모노 채널로 녹음되어 있다.

모델의 입력으로는 80차의 log-mel spectrogram을 사용하였다. 이때 윈도우의 크기는 25 ms이며 10 ms

Table 3. Output set of grapheme-based conformer.

cardinal vowel	ㅏ	ㅑ	ㅓ	ㅕ	ㅗ	ㅛ	ㅜ	ㅠ	ㅡ
	ㅣ								
cardinal constant	ㄱ	ㄴ	ㄷ	ㄹ	ㅁ	ㅂ	ㅅ	ㅇ	ㅈ
	ㅊ	ㅋ	ㅌ	ㅍ	ㅎ				
dobule constant	ㄲ	ㄸ	ㅃ	ㅆ	ㅉ				
compound constant	ㄳ	ㄵ	ㄶ	ㄺ	ㄻ	ㄼ	ㄽ	ㄾ	ㄿ
	ㅄ	ㅀ							
double vowel	ㅐ	ㅒ	ㅔ	ㅖ	ㅘ	ㅙ	ㅚ	ㅝ	ㅞ
	ㅟ	ㅢ							

Table 4. Output set of phoneme-based conformer.

onset	p0:ㅂ	ph:ㅍ	pp:ㅃ	t0:ㄷ	th:ㅌ	tt:ㄸ
	k0:ㄱ	kh:ㅋ	ss:ㅆ	h0:ㅎ	c0:ㅈ	cc:ㅉ
	mm:ㅁ	nn:ㄴ	rr:ㄹ	kk:ㄲ	s0:ㅅ	
coda	pf:ㅂ	tf:ㄷ	th:ㅌ	kf:ㄱ	kh:ㅋ	kk:ㄲ
	s0:ㅅ	ss:ㅆ	ch:ㅊ	mf:ㅁ	nf:ㄴ	ng:ㅇ
	ll:ㄹ	ks:ㄳ	nc:ㄵ	nh:ㄶ	lb:ㄼ	ls:ㄽ
	lt:ㄾ	lp:ㅀ	lh:ㄶ	ps:ㅄ	h0:ㅎ	c0:ㅈ
	lk:ㄺ	lm:ㄻ				
mono-phthong	ii:ㅣ	ee:ㅔ	qq:ㅐ	aa:ㅏ	xx:ㅡ	vv:ㅓ
	uu:ㅜ	oo:ㅗ	ya:ㅑ	yq:ㅒ	yv:ㅕ	yu:ㅠ
	yo:ㅛ	wi:ㅟ	wo:ㅚ	wq:ㅙ	wv:ㅝ	xi:ㅢ
	ye:ㅖ	yq:ㅒ	we:ㅞ	wa:ㅚ		

윈도우 시프트를 가진다.

모델의 출력은 자소와 음소로 달리 하여 2개의 모델을 학습하였다. 한국어의 경우, 음절은 초성, 중성, 종성으로 구성되어 있고 초성의 경우가 19개, 중성의 경우가 21개, 종성의 경우가 27개로 이론상 11,172개의 조합이 가능한다. 따라서 음절 기반 모델을 수렴 시키는 것은 매우 어렵기 때문에 자소 기반과 음소 기반을 비교하였다. 자소 기반 모델의 출력 집합은 기본 모음 10개, 기본 자음 14개, 쌍자음 5개, 복자음 11개, 복모음 11개로 <blank>와 띄어쓰기를 포함한 53개의 출력 노드 개수를 가진다. 음소 기반 모델은 G2P에 정의된 음소 집합을 사용하며 마찬가지로 <blank>와 띄어쓰기를 포함한 총 58개의 출력 집합을 사용하였다. Tables 3과 4는 각각 자소 기반 모델과 음소 기반 모델의 출력 집합을 나타낸다.

실험에 사용한 모델은 RNN-T 기반의 conformer를 사용하였다. 구현은 tensorflow 2.0을 사용하여 구현

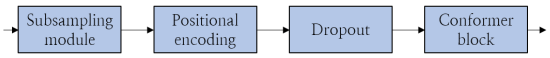

Fig. 3. (Color available online) Structure of conformer encoder.

Fig. 4. (Color available online) Structure of feed-forward module.

Fig. 5. (Color available online) Structure of multi-head self attention module.

하였다. Conformer의 예측 네트워크는 1층의 LSTM을 사용하였으며 전사 네트워크는 conformer 인코더를 사용하였다. Conformer 인코더는 16개의 conformer 블록으로 구성되었다. Fig. 3은 conformer 인코더의 구조를 나타낸다.

Conformer 블록은 feedforward 모듈과 multi-head self attention 모듈, 콘볼루션 모듈로 구성되며 Figs. 4~6은 각각 feedforward 모듈, multi-head self attention 모듈, convolution 모듈을 보여준다.

multi head self attention은 36차원을 가지는 4개의 헤드를 사용하였으며 feedforward 모듈은 최종 dropout을 거친 결과에 0.5를 곱한 후 residual 경로를 적용한다.

음소 기반 모델의 디코딩 방식은 end-to-end 모델의 출력 확률을 토대로 U transducer를 만든 후 기존의 L transducer와 conposition한 후 최단 경로를 결과로 반환한다. L transducer를 만드는데 필요한 단어는 학습자료에 등장하는 단어 중 빈도수로 단어의 개수를 바꿔가며 실험하였으며 그에 대한 음소열을 생성하기 위한 G2P는 규칙 기반의 한국어 G2P를 사용하였다. WFST를 만들고 compositon 및 최단 경로를 구하는 연산은 openfst를 사용하였다. 자소 기반 conformer의 디코딩 방식은 빔의 크기를 5로 하여 실험하였다. 실험결과 자소 기반 모델은 3.8%의 CER을 보였으며 음소 기반 모델에 L transducer를 적용한 모델은 3.4%의 성능을 보여 약간의 성능 향상을 보였다. Table 5는 L transducer의 단어 개수에 따른 성능 변화를 나타낸다.

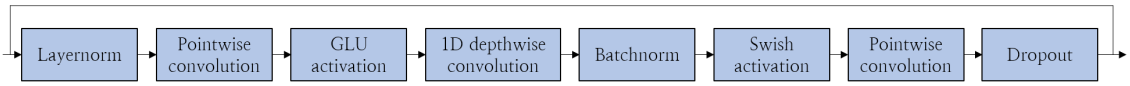

Fig. 6. (Color available online) Structure of convolution module.

Table 5. CER according to word count of L transducer.

Model	Size of vocab	CER (%)
grapheme-based conformer + beam search	-	3.80 %
phoneme-based conformer + WFST	80 K	5.45 %
	100 K	4.84 %
	128 K	3.40 %

V. 결 론

본 논문에서는 end-to-end 음성인식 방법에서 기존의 디코딩 방법으로 인해 발생하는 문제를 lexicon transducer를 활용하여 탐색 공간을 제한하는 방법을 제안하였다. 128,000개의 단어에 대해 한국어 G2P를 적용하여 L transducer를 만들었으며 U transducer와 L transducer를 composition하여 최소 코스트를 가지는 결과를 출력하는 디코딩 방식을 사용하였다. 실험은 현재 end-to-end 모델 중 가장 성능이 좋다고 평가받는 conformer를 사용하였으며 출력단을 음소 기반과 자소 기반으로 나누어 학습하였다. 음소 기반 모델에 제안하는 방법을 적용하여 디코딩한 모델과 자소 기반의 빔 탐색을 비교 실험 하였다. 약 270 h의 한국어 음성 코퍼스로 학습한 결과 테스트 셋에 대해 음소 기반 모델은 3.40 %의 음절 오류율을 보였으며 자소 기반 모델은 3.80 %의 성능을 보였다. 추후 연구에서는 언어모델 정보를 G transducer와 L transducer를 결합하여 언어모델이 적용된 결과를 비교 평가하고 여러 end-to-end 모델에 적용해 볼 계획이다.

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. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” Proc. NIPS. 1-11 (2017).
2. A. Gulati, J. Qin, C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, “Conformer: Convolution-augmented transformer for speech recognition,” Proc. Interspeech, 25-29 (2020).
3. M. Mohri, F. Pereira, and M. Riley, “Weighted finite-state transducer in speech recognition,” Computer Speech & Language, **16**, 69-88 (2002).
4. A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: Labeling unsegmented sequence data with recurrent neural networks,” Proc. ICML. 369-376 (2006).
5. A. Graves, “Sequence transduction with neural networks,” Proc. ICML. 1-9 (2012).
6. H. Park, S. Seo, M. Lim, D. Lee, Y. Kang, J. Oh, and J.-H. Kim, “Implementation of Korean grapheme-to-phoneme rules with morpheme analysis,” Proc. EECS. (2018).

저자 약력

▸ 손 현 수 (Hyunsoo Son)

2013년 2월 ~ 2019년 8월 : 서강대학교 컴퓨터공학과 학사
2019년 9월 ~ 현재 : 서강대학교 컴퓨터공학과 석사과정

▸ 박 호 성 (Hosung Park)

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

▸ 김 규 진 (Gyujin Kim)

2014년 2월 ~ 2019년 8월 : 서강대학교 컴퓨터공학과 학사
2019년 9월 ~ 현재 : 서강대학교 컴퓨터공학과 석사과정

▸ 조 은 수 (Eunsoo Cho)

2016년 3월 ~ 2021년 2월 : 서강대학교 중국문화/아트&테크놀로지/융합소프트웨어(인문콘텐츠융합) 학사
2021년 3월 ~ 현재 : 서강대학교 컴퓨터공학과 석사과정

▸ 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