한국어 자모단위 음성인식 결과 후보정을 위한 신경망 기반 자모 병합 방법론

임솔이*○1, 이원준*○2, 이근배†1,2, 김윤수†1,2
포항공과대학교 인공지능대학원 [1]
포항공과대학교 컴퓨터공학과 [2]
{solee0022, lee1jun, gary, yunsu.kim}@postech.ac.kr
*공동 1저자

Enhancing Korean Alphabet Unit Speech Recognition with Neural Network-Based Alphabet Merging Methodology

Solee Im*○1, Wonjun Lee*○2, Gary Geunbae Lee†1,2, Yunsu Kim†1,2
POSTECH, Graduate School of Artificial Intelligence[1], Computer Science and Engineering[2]
*Equal Contribution

요 약

이 논문은 한국어 음성인식 성능을 개선하고자 기존 음성인식 과정을 자모단위 음성인식 모델과 신경망 기반 자모 병합 모델 총 두 단계로 구성하였다. 한국어는 조합어 특성상 음성 인식에 필요한 음절 단위가 약 2900자에 이른다. 이는 학습 데이터셋에 자주 등장하지 않는 음절에 대해서 음성인식 성능을 저하시키고, 학습 비용을 높이는 단점이 있다. 이를 개선하고자 음절 단위의 인식이 아닌 51가지 자모 단위(ㄱ-ㅎ, ㅏ-ㅙ)의 음성인식을 수행한 후 자모 단위 인식 결과를 음절단위의 한글로 병합하는 과정을 수행할 수 있다[1]. 자모단위 인식결과는 초성, 중성, 종성을 고려하면 규칙 기반의 병합이 가능하다. 하지만 음성인식 결과에 잘못인식된 자모가 포함되어 있다면 최종 병합 결과에 오류를 생성하고 만다. 이를 해결하고자 신경망 기반의 자모 병합 모델을 제시한다. 자모 병합 모델은 분리되어 있는 자모단위의 입력을 완성된 한글 문장으로 변환하는 작업을 수행하고, 이 과정에서 음성인식 결과로 잘못인식된 자모에 대해서도 올바른 한글 문장으로 변환하는 오류 수정이 가능하다. 본 연구는 한국어 음성인식 말뭉치 KsponSpeech를 활용하여 실험을 진행하였고, 음성인식 모델로 Wav2Vec2.0 모델을 활용하였다. 기존 규칙 기반의 자모 병합 방법에 비해 제시하는 자모 병합 모델이 상대적 음절단위오류율(Character Error Rate, CER) 17.2% 와 단어단위오류율(Word Error Rate, WER) 13.1% 향상을 확인할 수 있었다. [1]

주제어: 음성인식, 오류수정, 한국어, 자모인식

1. 서론

한국어 자모 단위 음성인식 모델은 기존 한국어 음성인식 모델들이 예측 토큰을 음절 2904자로 둔 것과 달리 자모 51자로 두어 음성인식 성능을 향상시킬 수 있었다. 하지만 자모 단위의 시퀀스를 음절 단위의 문장으로 바꾸어 주는 과정에서 잘못 인식된 자모가 포함되어 있어도 오류 교정없이 규칙 기반으로 병합하는 한계가 있었다. 이러한 이유로 본 논문에서는 한국어 자모 단위 음성인식 모델이 일부 자모를 잘못 인식하더라도 올바른 음절 단위 문장으로 병합해주는 모델을 만들어 문제를 해결하고자 한다. 본 연구는 자모 단위의 음성인식을 수행한 후 해당 자모 시퀀스를 음절 단위 문장으로 변환해주는 두 단계로 진행된다. 기존 자모 단위 음성인식 연구에서 후처리 과정을 추가하여 최종 음성인식 성능을 높이는 것을 목표로 한다.

2. 관련 연구

최근 한국어 음성인식에 대한 다양한 연구가 이루어지고 있다. 2904개의 음절을 토큰으로 사용하는 Listen, Attend and Spell(LAS)[2]을 한국어 데이터셋 KsponSpeech[3]로 학습시킨 연구[4]와 Wav2Vec2.0으로 한국어 데이터셋을 학습시킨 연구가 있다[5]. [1]에서는 Connectionist Temporal Classification(CTC)[6] 디코더와 디코더 인식값을 51자의 자모로 두어 한국어 음성인식에서 좋은 성능을 올릴 수 있었다. CTC는 정렬 정보 없이 음성 입력값에 대한 출력값을 바로 End-to-End로 얻을 수 있어 많은 음성인식 연구에서 CTC 디코더를 사용하고 있다[7, 8, 9]. **자모 단위 CTC 디코더** 한국어 음성인식 모델의 출력값을 음절이 아닌 한글의 자모 51가지로 두고 음성인식을 수행한다. 기존 다른 한국어 음성인식 모델들[2, 4, 5]은 2904가지으로 이루어진 한글 음절 단위를 예측했다. [1]은 자모 단위로 음성을 인식하고 규칙 기반으로 자모를 병합했다. 이 연구는 신경망 기반 자모 병합 모델을 활용하여 음성인식 성능을 높이고자 한다. 음성인식의 정확도를 높이기 위해 음성인식 모델 자체의 성능을 향상시키는 방법도 있지만 음성인식 결과에 대한 오류 교정을 통해서 추가로 정확도를 높일 수 있다. 음성인식 후처리란 결과값 대해 맞춤법, 문법, 띄어쓰기 오류를 교정하는 것이다. 추가적인 음성 데이터셋 없이 텍스트 데이터만으로도 음성인식 성능을 높일 수 있기 때문에 후처리에 대한 연구가 활발히 진행되고 있다. 최근에는 Transformers 기반의 인코더-디코더 모델을 활용한 후처리 연

구가 제시되었다[10, 11, 12, 13, 14]. 한국어 음성인식 후처리 연구는 LSTM 모델을 기반으로 한 음절 단위 단어 교정이 있었다[15]. 본 연구에서는 Transformenrs 기반의 인코더-디코더를 이용해 자모 단위 음성인식 결과를 병합 및 후처리 하는 방법을 제시한다.

3. 본론

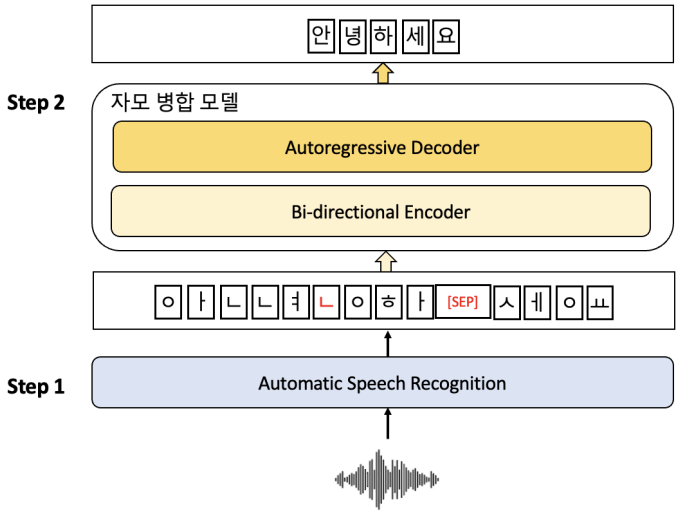

그림 1. **자모 병합 과정 순서도: Step1.** 음성 데이터를 입력값으로 받아 자모 단위 음성인식 진행 **Step2.** 음성인식 결과값으로 나온 자모 시퀀스를 입력값으로 받아 음절 단위 문장으로 변환

연구는 크게 두 단계로 진행되었다1. 첫 번째는 자모 단위 음성인식 모델이 음성신호를 입력받아 자모 단위 시퀀스를 출력한다. 두 번째는 자모 병합 모델이 자모 단위 시퀀스를 입력받아 음절 단위 문장을 출력한다.

3.1 자모 단위 음성인식 모델

[1]의 자모 단위 음성인식 모델 아키텍쳐를 사용했다. [1]이 영어와 한국어 두 개의 언어를 인식하는 다국어 음성인식 모델을 만들기 위해 CTC 언어사전에 영어 28자를 넣어주었던 것과 달리 CTC가 한국어 자모 중 하나로 음성에 대한 텍스트 라벨을 예측할 수 있도록 자음과 모음 51자(ㄱ-ㅎ, ㅏ-ㅙ)로 구성하였다.

3.2 자모 병합 모델

3.2.1 규칙 기반 자모 병합

[1]이 자모 단위 음성인식 결과를 병합해준 방식이다. 자모 시퀀스를 한글의 초성-중성-종성 규칙에 따라 자모를 병합해주는 것이다. 음성에 대한 자모 시퀀스가 항상 정확하게 만들어진다면 규칙 기반으로 자모를 결합해주어도 언제나 정확한 음절 단위 문장을 얻을 수 있다. 하지만 음성인식 출력값에 예측 오류 자모가 포함된다면 최종 병합 결과로 잘못된 단어로 이루어진 문장이 만들어진다는 문제가 있다.

3.2.2 신경망 모델

단순 규칙 기반의 자모 결합이 아니라 한국어 텍스트 데이터로 학습된 자모 병합 모델을 사용해 자모 시퀀스를 한국어 음절 문장으로 변환하고자 한다. 모델은 Transformers 기반의 인코더-디코더 모델인 mBART의 아키텍쳐를 가져와 사용했다. 규칙 기반 변환과 달리 음성인식 모델이 잘못 예측한 자모 오류를 교정해주면서 음절 문장으로 변환해주는 후처리 모델이다. 모델을 학습시키기 위해 한국어 텍스트 데이터셋을 구축하였다2. 노이즈 데이터는 음성인식 모델을 사용해 후술할 방법3.2.3으로 증강해주었다.

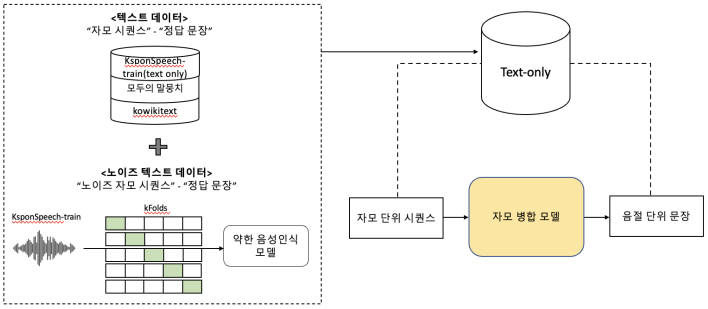

그림 2. 자모 병합 모델 데이터셋. 텍스트 데이터 및 노이즈 텍스트 데이터 생성(좌), 신경망 기반 자모 병합 모델 학습(우)

3.2.3 노이즈 데이터 증강

자모 병합 모델이 노이즈 자모 시퀀스를 학습할 수 있도록 약한 음성인식 모델을 활용해 노이즈 데이터를 생성해주었다. 약한 음성인식 모델이란 음성인식 모델 학습 중 평가 데이터셋에 대한 평가 지표가 WER=0.3 이상인 모델 체크포인트를 칭한다. 무작위로 틀린 자모를 시퀀스에 넣거나 지워서 노이즈한 데이터를 생성하는 방식이 아니라 실제로 음성인식 모델이 흔히 만들어 내는 예측 실수를 노이즈 데이터로 만들어 주기 위해서 이다. 노이즈 데이터 생성을 위한 약한 음성인식 모델을 학습시키기 위해서 kFold 학습 방식을 사용하였다. 학습 데이터를 5분할 하여 4개의 세트로 음성인식 모델을 학습시키고 남은 1개의 세트를 약한 음성인식 모델에 넣어 노이즈 데이터를 생성하였다. 이 과정을 5번 진행해주었다.

4. 실험 및 연구결과

4.1 데이터셋

음성인식 모델 학습을 위한 데이터로는 AI HUB 한국어 음성 데이터셋 KsponSpeech[3]를 사용하였다. 자음-한글 변환 모

델 학습 데이터로는 KsponSpeech 학습 텍스트 데이터, 모두의 말뭉치 국립국어원 일상 대화 음성 말뭉치 2020 - 한국어 자유 발화 텍스트 데이터, kowikitext 그리고 KsponSpeech 오디오 데이터로 만든 노이즈 텍스트 데이터를 사용했다. 자세한 데이터셋의 정보는 표1에서 확인할 수 있다.

데이터셋	KsponSpeech	모두의말뭉치	Kowiki
텍스트	600,000 문장	870,162 문장	14,528,095 문장
오디오	1000 시간	-	-

표 1. 실험에 활용한 데이터셋 정보. 음성인식 모델 학습에는 KsponSpeech 데이터셋의 오디오 데이터를 사용하였고, 자모 병합 모델 학습에는 세 데이터셋의 텍스트 데이터와 KsponSpeech 노이즈 텍스트 데이터를 사용하였다.

4.2 학습 방법

4.2.1 자모 단위 음성인식 모델

Wav2Vec2.0 인코더와 자모 단위 CTC 디코더의 구조이다. Wav2Vec2.0을 학습시키기 위해 huggingface/transformers 라이브러리를 사용하였다[16]. 모델 학습을 위해 총 8대의 NVIDIA A100-80GB를 사용하였다. 각 GPU 당 Batch Size를 20으로 두고, gradient accumulation 값을 2로 두어 총 320(8*2*20)배치로 학습하였다. 모델은 30 Epochs으로 설정해주었고 learning rate는 3e-4, 옵티마이저는 AdamW optimzer[17]를 사용했다. 학습 과정 도중 중간 평가를 통해 dev셋에서 CER이 가장 낮은 모델을 최종 체크포인트로 사용했다.

4.2.2 자모 병합 모델

Pre-processing sentencepiece[18] 모델을 토크나이저로 사용하였다. 음절 단위 문장과 자모 단위 문장으로 각각 구성된 코퍼스를 사용하여 32,000개의 토큰으로 이루어진 sentencepiece 토크나이저를 학습하였다.

Training 보통 mBART 모델을 사용할 때 사전 학습된 Denoising Transformer 인코더-디코더 모델을 사용한다. 하지만 우리 연구의 경우 자모라는 새로운 단위의 문자를 입력값으로 받아 사용하기 때문에 사전 학습된 모델을 사용할 수 없었다. 그래서 mBART의 모델 아키텍쳐는 가져오되 사전 학습된 모델값은 사용하지 않았다. 그림3과 같이 “정답 자모 시퀀스 - 정답 음절 문장”와 “노이즈 자모 시퀀스 - 정답 음절 문장”으로 이루어진 텍스트 데이터셋을 가지고 자모 병합 모델을 학습시켰다. 음성인식 모델이 자모를 잘못 예측하더라도 정답 음절 문장을 출력값으로 얻을 수 있도록 하였다. 모델 학습 시 총 8대의 NVIDIA A100-80GB을 사용하였다. 각 GPU 당 Batch Size를 20으로 두고 gradient accumulation 값을 2로 두어 총 320(8*2*20)배치로 학습하였다. 모델은 10 Epochs으로 설정했고 learning rate는 3e-4, 옵티마이저는 AdamW optimzer를 사용했다. 음성인식 모델과 마찬가지로 중간 평가를 통해 dev셋에서 CER이 가장 낮은 모델을 최종 체크포인트로 사용하였다.

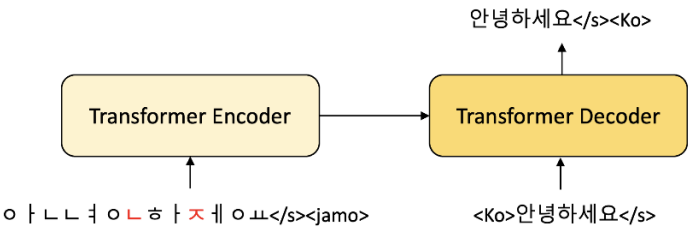

그림 3. 자모 병합 모델

4.3 실험결과

성능지표로는 CER과 WER을 사용하였다. 자모 음성인식 결과를 규칙 기반으로 합쳐준 것을 baseline으로 두고 이번 연구에서 학습한 자모 병합 모델로 후처리 해준 것과 성능 비교 하였다2. KsponSpeech DEV 데이터셋에 대해서 CER은 24.8% WER 20.2% 성능이 향상되었다. EVAL-CLEAN 데이터셋에 대해서는 CER은 17.2%, WER은 13.1% 성능이 향상되었다. 마지막으로 EVAL-OTHER 데이터셋에 대해서는 CER은 19.9%, WER은 14.5% 성능이 향상되었다. 모든 경우 규칙 기반으로 자모 결과값을 병합해준 것보다 자모 병합 모델을 사용한 것이 성능이 향상되었다.

표 3과 같이 자모 음성인식 결과에 잘못 인식된 경우가 있어도 규칙 기반과 다르게 오류를 수정하면서 자모 병합이 가능함을 확인할 수 있었다. 예를 들어 자모 음성인식 모델이 ‘ㄷㅗㅜㅁ’과 같이 자모를 잘못 인식했을 때 규칙 기반은 오류 수정없이 ‘도ㅜㅁ’이라고 병합했지만 자모 병합 모델은 ‘도움’과 같이 오류 수정 후 병합됨을 확인할 수 있었다.

5. 결론

본 연구에서 신경망 기반 자모 병합 모델을 이용한 후처리 기법으로 추가적인 한국어 음성 데이터 없이 텍스트 데이터만을 이용해 한국어 음성인식 성능을 향상시켰다. KsponSpeech 데이터셋에서 규칙 기반 자모 병합보다 CER은 17.2%, WER은 13.1% 향상을 보였다. 자모 단위 한국어 음성인식 모델이 단순 규칙 기반으로 자모를 조합한다는 한계점에서 출발한 연구가 몇 개의 자모 오류 교정만으로도 음성인식의 성능을 향상시킬 수 있음을 확인할 수 있었다. 자모 오류 교정뿐만 아니라 띄어쓰기와 같은 문법적 오류도 잡아줘 문장 전체의 정확도를

	DEV		EVAL$_C$*LEAN*		EVAL$_O$*THER*	
MODEL	CER	WER	CER	WER	CER	WER
규칙 기반 (baseline)	12.9664	27.8700	11.8689	28.2419	13.7392	34.4388
자모 병합 모델 (proposed)	**9.7510**	**22.2239**	**9.8266**	**24.5784**	**10.9965**	**29.4548**

표 2. 한국어(KsponSpeech) 음성인식 성능 지표

(1)	INPUT: ㅇ ㅐㄷㅡㄹㅎ ㅏㄴㅌ ㅔ ㄷㅗㅜㅁ ㅂ ㅏㄷㅇ ㅏㅆㅈ ㅣ
	규칙 기반: 애들한테 도ㅜㅁ 받았지 자모 병합 모델: 애들한테 도움 받았지
(2)	INPUT: ㅇ ㅏㄴ ㅐㄱ ㅏㅈㅗㅁ ㄴ ㅏㅈㄷ ㅏ
	규칙 기반: 아 내가 좀 낮다 자모 병합 모델: 아 내가 좀 낫다
(3)	INPUT: ㅇ ㅏㅁㅜㄹ ㅐㄷㅗ ㅇ ㅣㅈ ㅔ ㅎ ㅏㄴㅅ ㅏㄹ ㅎ ㅏㄴㅅ ㅏㄹㅇ ㅣ
	규칙 기반: 아무래도 이제 한살 한살이 자모 병합 모델: 아무래도 이제 한 살 한 살이

표 3. 한국어(KsponSpeech) 음성인식 후처리. (1)발음 인식 오류 - 모델 자체 문제 (2)발음 인식 오류 - 발음의 차이가 없는 단어 (3)띄어쓰기 오류

높일 수 있었다. 더 나아가 특정 도메인의 대화를 음성인식 할 경우 특정 도메인 관련 텍스트 데이터로 자모 병합 모델을 추가 학습시켜 준다면 음성인식에서 흔히 발생하는 OOD(Out-Of-Domain) 문제도 해결할 수 있을 것이라 기대한다.

감사의 글

이 연구는 2021년도 산업통상자원부 및 산업기술평가관리원(KEIT) 연구비 지원에 의한 연구임(20015007). 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터육성지원사업의 연구결과로 수행되었음(IITP-2023-2020-0-01789).

참고문헌

[1] 이원준, "다국어 음성인식을 위한 언어별 출력 계층 구조 wav2vec2.0," 2021. [Online]. Available: https://koreascience.kr/article/CFKO202130060715833.pdf

[2] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell," 2015.

[3] J.-U. Bang, S. Yun, S.-H. Kim, M.-Y. Choi, M.-K. Lee, Y.-J. Kim, D.-H. Kim, J. Park, Y.-J. Lee, and S.-H. Kim, "Ksponspeech: Korean spontaneous speech corpus for automatic speech recognition," *Applied Sciences*, Vol. 10, No. 19, 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/19/6936

[4] S. Kim, S. Bae, and C. Won, "Kospeech: Open-source toolkit for end-to-end korean speech recognition," 2020.

[5] J. Kim and P. Kang, "K-wav2vec 2.0: Automatic speech recognition based on joint decoding of graphemes and syllables," 2021.

[6] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06, p. 369–376, 2006. [Online]. Available: https://doi.org/10.1145/1143844.1143891

[7] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.

[8] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," 2021.

[9] A. Gulati, J. Qin, C.-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," 2020.

[10] S. Dutta, S. Jain, A. Maheshwari, G. Ramakrishnan, and P. Jyothi, "Error correction in ASR using sequence-to-sequence models," *CoRR*, Vol. abs/2202.01157, 2022. [Online]. Available: https://arxiv.org/abs/2202.01157

[11] L. Mai and J. Carson-Berndsen, "Unsupervised domain adaptation for speech recognition with unsupervised error correction," 2022.

[12] Y. Zhao, X. Yang, J. Wang, Y. Gao, C. Yan, and Y. Zhou, "BART based semantic correction for mandarin automatic speech recognition system," *Interspeech 2021*, aug 2021. [Online]. Available: https://doi.org/10.21437%2Finterspeech.2021-739

[13] C.-H. Kuo and K.-Y. Chen, “Correcting, rescoring and matching: An n-best list selection framework for speech recognition,” *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 729–734, 2022.

[14] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, “Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” 2019.

[15] 진혜원, 이아현, 채예진, 박수현, 강유진, and 이수원, “Lstm 기반의 seq2seq 모델을 이용한 한국어 음성인식 오류 교정 방법,” *한국컴퓨터정보학회논문지*, Vol. 26, No. 10, pp. 1–7, 2021.

[16] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, “Transformers: State-of-the-art natural language processing,” *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Oct. 2020. [Online]. Available: https://aclanthology.org/2020.emnlp-demos.6

[17] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” 2019.

[18] T. Kudo and J. Richardson, “Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing,” 2018.