Deep CNN 기반의 한국어 음소 인식 모델 연구

홍윤석*, 기경서*, 권가진*
*서울대학교 융합과학부
e-mail : yshong93@snu.ac.kr, kskee88@snu.ac.kr, ggweon@snu.ac.kr

Korean Phoneme Recognition Model with Deep CNN

Yoon Seok Hong*, Kyung Seo Ki *, Gahgene Gweon *
*Dept. of Transdisciplinary Studies, Seoul National University

요 약

본 연구에서는 심층 합성곱 신경망(Deep CNN)과 Connectionist Temporal Classification (CTC) 알고리즘을 사용하여 강제정렬(force-alignment)이 이루어진 코퍼스 없이도 학습이 가능한 음소 인식 모델을 제안한다. 최근 해외에서는 순환 신경망(RNN)과 CTC 알고리즘을 사용한 딥 러닝 기반의 음소 인식 모델이 활발히 연구되고 있다. 하지만 한국어 음소 인식에는 HMM-GMM 이나 인공 신경망과 HMM 을 결합한 하이브리드 시스템이 주로 사용되어 왔으며, 이 방법은 최근의 해외 연구 사례들보다 성능 개선의 여지가 적고 전문가가 제작한 강제정렬 코퍼스 없이는 학습이 불가능하다는 단점이 있다. 또한 RNN 은 학습 데이터가 많이 필요하고 학습이 까다롭다는 단점이 있어, 코퍼스가 부족하고 기반 연구가 활발하게 이루어지지 않은 한국어의 경우 사용에 제약이 있다. 이에 본 연구에서는 강제정렬 코퍼스를 필요로 하지 않는 CTC 알고리즘을 도입함과 동시에, RNN 에 비해 더 학습 속도가 빠르고 더 적은 데이터로도 학습이 가능한 합성곱 신경망(CNN)을 사용하여 딥 러닝 모델을 구축하여 한국어 음소 인식을 수행하여 보고자 하였다. 이 모델을 통해 본 연구에서는 한국어에 존재하는 49 가지의 음소를 추출하는 세 종류의 음소 인식기를 제작하였으며, 최종적으로 선정된 음소 인식 모델의 PER(Phoneme Error Rate)은 9.44 로 나타났다. 선행 연구 사례와 간접적으로 비교하였을 때, 이 결과는 제안하는 모델이 기존 연구 사례와 대등하거나 조금 더 나은 성능을 보인다고 할 수 있다.

1. 서론

음소 인식은 사람의 음성 데이터로부터 발음의 기본 단위라고 할 수 있는 음소(phoneme)를 식별하여 구분해 주는 작업으로, 음성 인식(Automation Speech Recognition) 분야에서의 주요한 요소 기술이다. 기존에는 음소 인식을 수행하기 위한 주요 모델로 MFCC(Mel-Frequency Cepstrum Coefficient)를 사용한 은닉 마르코프 모델(HMM)과 가우시안 믹스처 모델(GMM) 과 같은 확률 함수 모델을 사용하여 음소 인식 모델을 구성하는 것이 일반적이었다[1, 2, 3, 4]. 하지만 최근 심층 인공신경망을 도입한 음소 인식 모델이 점차 연구되면서, 딥 러닝 기반의 모델을 통해 음소 인식을 수행함으로써 더 나은 성능을 내고자 하는 시도가 계속해서 이루어지고 있다[5, 6, 7].

그러나 한국어 음소 인식에 대한 연구는 해외의 연구 상황에 비해 매우 미미한 상황이다. 최근의 한국어 음소 인식에 대한 연구를 살펴보면, 대부분이 HMM, GMM 을 사용하여 모델을 구성한 연구가 대다수이며, 딥 러닝을 도입한 연구는 찾아보기 어렵다. 또한 기존의 음소 인식기 개발을 위해서는 전문가에 의해 제작된 강제정렬(force-alignment) 코퍼스가 반드시 필요했는데, 한국어에서는 공개되어 있는 강제정렬된 코퍼스가 매우 부족한 실정이다. 따라서 본 연구에서는 (1) 해외 사례에서 보고된 바 있는 성능 개선 효과를 기대하며 딥 러닝 기법을 도입하여 합성곱 신경망(CNN)기반의 음소 인식기를 개발하였으며, (2) 강제정렬 없이도 학습이 가능할 수 있도록 제안한 음소 인식 모델에 CTC-알고리즘을 적용하여 보고자 하였다.

본 논문의 구성은 다음과 같다. 우선 2 장에서는 음소 인식 태스크를 수행한 선행 연구 사례에 대해 논의한다. 3 장에서는 본 연구를 수행하기 위해 적용한 합성곱 신경망(CNN) 모델 및 Connection Temporal Classification(CTC) 디코딩에 대해 설명한다. 마지막으로 4 장에서는 실험 결과를 보고하며, 본 연구에서 제안한 모델의 적용 가능성 및 향후 연구 방향을 논의한다.

2. 관련 연구

2-1 인공 신경망 기반의 음소 인식

인공 신경망을 사용한 음소 인식은 해외에서 30 년이 넘는 기간 동안 활발하게 진행되어 왔다. 특히 인공신경망을 활용한 음소 인식은 1990 년부터 연구되어 왔으며[5], 심층 신경망(DNN)과 HMM 을 결합한 형태의 DNN-HMM 하이브리드 모델도 제안되어

HMM 만 사용한 모델보다 높은 인식률을 보인 것으로 보고되고 있다[8, 9]. 가장 활발한 연구 결과가 보고되고 있는 영어의 경우, HMM 을 사용한 음소 인식기의 Phoneme Error Rate (PER)은 24.25 였으며[3], 하이브리드 모델의 경우 20.25 의 결과가 나타났다[8]. 또한 최근에 들어서는 DNN 뿐만 아니라 시계열 정보에 더 적합한 것으로 알려져 있는 순환 신경망(RNN)을 사용하여, PER 을 17.7 까지 낮추어 당대 최고 성능을 보였던 사례도 보고되고 있다[10]. 위의 연구들은 영어 음소 스피치 코퍼스인 TIMIT 을 사용하였으며 한국어로 된 음소 인식 결과는 찾아보기 어려웠다.

하지만 기존의 HMM-GMM 모델은 학습을 위해 오디오 데이터의 정보 손실이 발생하는 Mel-Frequency Cepstrum Coefficient(MFCC)와 같은 변환 Feature 를 사용하기 때문에 오디오 데이터를 그대로 사용하는 최근의 방법에 비해서는 성능 개선에 어려움이 있는 것으로 알려져 있으며, 순환 신경망은 시간 데이터에 대해 현재의 상태를 업데이트 하는 방식으로 학습이 진행되기 때문에 학습 속도가 느리고, 다른 구조에 비해 데이터 요구량이 더 많아서 학습에서 많은 제약이 존재한다. 특히 데이터가 많이 요구된다는 제약 때문에 음성인식 분야에서는 최근 들어 학습이 비교적 빠르고 쉬운 CNN 을 활용하고자 하는 연구가 많이 이루어지고 있다[11, 12, 13, 14].

2-2 CTC 알고리즘

CTC 는 강제 정렬되지 않은 데이터를 사용하여 학습을 시킬 수 있도록 고안된 알고리즘이다[15]. 기존의 음소 인식에서 많이 사용하던 하이브리드 모델(DNN-HMM/GMM 모델)이나 순환 신경망 기반의 RNN / LSTM 모델은 강제 정렬된 오디오 코퍼스가 있어야 학습이 가능하다. 하지만 강제 정렬된 데이터는 만드는데 오랜 시간이 들기 때문에 데이터를 수집하는 데 어려움이 컸다. CTC 를 사용하게 되면 학습 과정에서 나뉘어진 타임 레이블 각각에 대한 음소 예측이 가능해지므로, 기존의 학습 방법처럼 미리 강제 정렬된 정보를 제공할 필요가 없어진다. 이에 CTC 의 등장으로 강제 정렬된 데이터 없이도 학습이 가능하게 되었으며, CTC 알고리즘을 음성 인식 분야에서 활용하는 연구들이 있다. 기존에 RNN 과 함께 사용하던 CTC 알고리즘을 CNN 과 결합하여 사용한 Palaz 외 (2015)의 연구 사례도 있으며[12], Hori 외 (2017) 는 CTC 알고리즘에 어텐션 메커니즘(Attention Mechanism)을 합친 CTC-attention 에 대해 연구를 진행하였다[16].

CTC 의 구조는 인풋 신호 X 에 대해 시퀀스 X 로부터 체인 룰(Chain Rule) 로 연결된 시퀀스 Y 을 찾는 형태이며, 수식은 다음과 같다.

$$P(Y|X) = \prod_i P(y_i \mid X, y_{<i})$$

CTC 는 인풋 신호 시퀀스 X 와 길이가 같은 시간 레이블(Temporal Label) 을 가진다. 이때, Y 의 레이블의 유형을 K 개라 가정하면, 한 개의 공백 레이블을 더 해진 (K+1)개의 유형이 시간 레이블로 예측된다. CTC-디코딩은 시간 레이블 시퀀스를 음소 시퀀스 Y 로 변환시켜주는 과정을 의미하며 CTC-디코딩 과정은 1) 시간 레이블에 포함되어 있는 중복 레이블을 제거한 뒤, 2) 공백 레이블을 제거하는 순서로 진행된다.

3. 실험 방법

본 연구에서는 음소 인식을 수행하기 위해 전사 자료와 음성 데이터만 있는 서울말 낭독체 코퍼스를 사용한다. Graph to Phoneme(G2P)를 사용하여 전사 자료를 음소 정답지를 변환 한 뒤, 3 가지 후보 모델로 학습을 실시하여 결과를 확인하는 방식으로 실험을 진행하였다. 구체적인 실험 과정은 아래와 같다.

3-1 실험 데이터

본 연구에서는 음소 인식기 학습을 위해 국립 국어연구원에서 만든 '서울말 낭독체' 코퍼스[17]를 사용하였다. 서울말 낭독체 코퍼스는 서울말을 사용하는 80 명(20 대~60 대 남, 여 각각 20 명) 의 음성 녹음 파일과 음소 강제 정렬이 되어 있지 않은 전사 파일로 구성되어 있다. 음성 파일은 총 71,216 개, 약 180 시간의 분량이다. 모두 16,000 Hz 로 샘플링 되어 있고, 노이즈가 없는 실험실 환경에서 한 문장씩 녹음되었다. 전사 자료는 총 19 개의 단편 소설 및 수필이며, 각각 대략 50 문장으로 구성되어 있다. 실험 전 음성 코퍼스를 모두 검수하였으며, 음성 파일과 전사 자료가 일치하지 않는 파일과 명확하지 않은 발음으로 녹음된 파일은 본 연구에서 사용하지 않았다. 학습을 시키기 위해 음성 코퍼스를 6:1:1 의 비율로 각각 Train, Validation, Test 로 무작위로 나누어서 사용하였다.

3-2 실험 환경

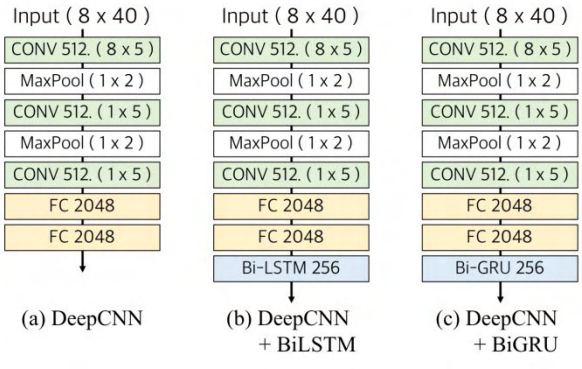
음소 인식기 모델을 구축하기 위하여 인풋 Feature 로 Mel-Spectrogram 을 사용하여 CNN 모델을 학습시키고, CTC-디코딩을 통해 최종적으로 음소 아웃풋을 얻어 정답지와 비교하여 평가를 수행하였다. Mel 주파수 대역폭은 40 으로 설정하였다. 프레임은 5ms 의 길이로 각각 나누었으며, 타임 레이블에서의 예측을 용이하게 하기 위하여 프레임 간의 겹침(overlapping) 은 없었다. 이렇게 하여 1 프레임당 1 개의 시간 레이블 C 가 추출되며, 시간 레이블 C 는 총 50 개의 유형이 있다. 이 유형들은 각각 49 개의 한국어 음소와 1 개의 공백 레이블(blank)이 있다.

$$C \in \{ AA, AX, CH, EH, EY, \ldots, blank \}.$$

한편 정답지를 구성하기 위해 전사 자료를 표준 발음에 맞춰 음소 시퀀스 Y* 로 변환하였으며, 변환에는 한국어 문장(Graph)를 표준 음소(Phoneme)로 변경해주는 KoG2P [18] 시스템을 사용하였다.

이번 실험에서는 그림 1 과 같이 3 가지의 모델 아키텍처를 구성하여 비교하였다. 비교에 활용된 모델

들은 (a) CNN 만 사용한 모델, (b) CNN 과 BiLSTM 레이어를 사용한 모델 그리고 (c) CNN 과 BiGRU 레이어를 사용한 모델이다.



(그림 1) VGG 를 변형한 DeepCNN, DeepCNN+BiLSTM, DeepCNN+BiGRU 음소 인식 모델 아키텍처

세 아키텍처에 공통으로 사용한 CNN 모델은 VGG16 네트워크를 간략하게 변형한 모델이다[13, 14]. VGG16 네트워크는 필터 사이즈가 작은 세 겹의 컨벌루션 레이어(Convolution Layer)로 구성되어 있는 CNN 을 여러 겹으로 깊게 쌓는 모델로, 본 연구에서는 세 겹의 컨벌루션 레이어로 구성된 CNN 을 각각 1 층씩 쌓는 구조로 아키텍처를 구성하였으며, 각각의 CNN 컨벌루션 레이어마다 Max Pooling 을 진행하였다. (b), (c) 에서는 1 프레임마다 CNN 레이어를 통해 추출된 시퀀스를 인풋으로 사용하는 BiLSTM, BiGRU 레이어를 CNN 레이어 뒤에 추가하여, CNN 만을 사용한 모델에 비해 성능이 더 나은지를 확인하여 보고자 하였다. 활성 함수(Activation Function)로는 clipped rectified-linear unit (clipped-ReLU) 를 사용하였다. 수식은 아래와 같다.

$$\text{Clipped} - \text{ReLU: } g(z) = \min\{\max\{0, x\}, 20\}.$$

모델 훈련에서는 Adam 옵티마이저를 사용하였다. 오버피팅(overfitting)을 막기 위해 20% 드랍아웃(dropout) 을 사용하였으며, 오디오 데이터에 무작위로 화이트 노이즈와 핑크 노이즈를 추가하였다. 세 모델 모두 15 주기(epoch)에서 손실 값(loss)의 변화가 매우 작아, 조기 중단(Early Stopping)을 통해 학습을 마무리 하였다.

4. 결과 및 토의

본 연구에서는 음소 인식기의 정확도를 측정하기 위해 예측한 레이블 시퀀스 $\hat{Y}$와 정답지인 골드 스탠다드 Y 사이의 PER 을 구하여 음소 인식기의 최종 성능을 평가하였다[15]. PER 스코어는 테스트 셋 S 에 대하여 $S \subseteq (X, Y)$ 일 때, $\hat{y}$를 음소 인식기의 아웃풋으로 정의하면, 아래 수식을 통하여 구할 수 있다. 수식에서 ED 는 Edit Distance 를 의미한다. PER 스코어는 음소 레이블의 에러 정도를 의미하므로, PER 스코어가 낮을 수록 성능이 좋은 모델을 의미한다.

$$LER(S) = \frac{1}{S}\sum \frac{ED(y, \hat{y})}{len(y)}$$

각 세 가지 종류의 모델에 대해 PER 스코어를 구한 결과, 표 1 과 같은 결과를 얻었다.

<표 1> 음소 인식 모델 결과

Model	PER
DeepCNN	**9.44**
DeepCNN + BiLSTM	10.476
DeepCNN + BiGRU	10.467

8800 개의 테스트 셋을 대상으로 수행한 실험 결과, LSTM 이나 GRU 와 같은 순환 신경망을 하단에 붙인 모델에 비해 CNN 만을 사용한 모델이 가장 나은 PER 결과를 보이는 것을 확인할 수 있었다. 이러한 결과는 다른 실험과 테스트 셋의 종류 및 실험 방법 등이 다르기 때문에 직접적인 비교는 불가능하나, 한국어 음소 인식 연구 사례로 가장 최근 보고된 바 있는 나민수 외[19]의 음소 인식기 성능과 비슷한 결과를 보였다. 나민수 외의 연구 사례는 147,263 개의 테스트 셋을 대상으로 19,541 개의 불일치를 보여, 12 정도의 PER 값이 나온 것으로 보고되고 있다.

본 연구를 통해 기존의 HMM-GMM 방식이나 하이브리드 방식을 사용하지 않고 딥 러닝만을 사용해서도 음소 인식에서 기존의 음소 인식 결과와 비슷한 PER 수치를 얻을 수 있다는 것을 확인할 수 있었다. 또한, CTC 가 강제정렬 없이도 음소 인식 태스크를 수행할 수 있게 해 준다는 것이 낮은 PER 수치를 통해 입증되었다. 이러한 결과를 감안한다면, 음소 인식만이 아니라 강제정렬 코퍼스가 부족한 음성 인식에서의 다른 태스크에서도 딥 러닝과 CTC 의 활용이 해결책이 될 수도 있을 것으로 보인다.

한편 실험 과정에서 세 모델 모두 대부분의 오류는 발음 시간이 짧고 진폭이 작은 자음에서 발생하였으며, 모음에서는 오류가 더 적게 나타나는 것을 확인할 수 있었다. 향후 실험에서는 이를 고려하여 자음에서의 오류를 더 낮출 수 있는 모델을 개발할 수 있도록 하고자 한다.

참고문헌

[1] Gales, Mark JF. "Maximum likelihood linear transformations for HMM-based speech recognition." Computer speech & language 12.2 (1998): 75-98.

[2] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[3] Glass, James R. "A probabilistic framework for segment-based speech recognition." Computer Speech & Language 17.2-3 (2003): 137-152.

[4] Schwarz, Petr, Pavel Matějka, and Jan Černocký. "Towards lower error rates in phoneme recognition." International Conference on Text, Speech and Dialogue. Springer, Berlin, Heidelberg, 2004.

[5] Waibel, Alexander, et al. "Phoneme recognition using time-delay neural networks." Readings in speech recognition. 1990. 393-404.

[6] Bengio, Yoshua. "A connectionist approach to speech recognition." Advances in Pattern Recognition Systems Using Neural Network Technologies. 1993. 3-23.

[7] Mohamed, Abdel-rahman, George E. Dahl, and Geoffrey Hinton. "Acoustic modeling using deep belief networks." IEEE Transactions on Audio, Speech, and Language Processing 20.1 (2012): 14-22.

[8] Seltzer, Michael L., and Jasha Droppo. "Multi-task learning in deep neural networks for improved phoneme recognition." Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013.

[9] Graves, Alex, Navdeep Jaitly, and Abdel-rahman Mohamed. "Hybrid speech recognition with deep bidirectional LSTM." Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on. IEEE, 2013.

[10] Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton. "Speech recognition with deep recurrent neural networks." Acoustics, speech and signal processing (icassp), 2013 ieee international conference on. IEEE, 2013.

[11] Palaz, Dimitri, Ronan Collobert, and Mathew Magimai Doss. "End-to-end phoneme sequence recognition using convolutional neural networks." arXiv preprint arXiv:1312.2137 (2013).

[12] Palaz, Dimitri, Mathew Magimai Doss, and Ronan Collobert. "Convolutional neural networks-based continuous speech recognition using raw speech signal." Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. IEEE, 2015.

[13] Heck, Michael, et al. "Ensembles of Multi-scale VGG Acoustic Models." Proc. Interspeech 2017 (2017): 1616-1620.

[14] Zhang, Ying, et al. "Towards end-to-end speech recognition with deep convolutional neural networks." arXiv preprint arXiv:1701.02720 (2017).

[15] Graves, Alex, et al. "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks." Proceedings of the 23rd international conference on Machine learning. ACM, 2006.

[16] Hori, Takaaki, et al. "Advances in joint CTC-attention based end-to-end speech recognition with a deep CNN encoder and RNN-LM." arXiv preprint arXiv:1706.02737 (2017).

[17] 국립국어원, “서울말 낭독체 발화 말뭉치,” 2003. URL: https://ithub.korean.go.kr

[18] 조예진, Korean Grapheme-to-Phoneme Analyzer (KoG2P), 2017. GitHub repository: https://github.com/scarletcho/KoG2P

[19] 나민수, 정민화 (2016). 한국어 발음열 자동생성과 음성인식에 기반한 음성전사 지원 시스템. 한국음성학회 봄 학술대회 발표논문집, 131-132.