pISSN 2005-8063
eISSN 2586-5854
2019. 3. 31.
Vol.11 No.1
pp. 41-49

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2019.11.1.041

Visual analysis of attention-based end-to-end speech recognition

Seongmin Lim · Jahyun Goo · Hoirin Kim*

School of Electrical Engineering, KAIST, Daejeon, Korea

Abstract

An end-to-end speech recognition model consisting of a single integrated neural network model was recently proposed. The end-to-end model does not need several training steps, and its structure is easy to understand. However, it is difficult to understand how the model recognizes speech internally. In this paper, we visualized and analyzed the attention-based end-to-end model to elucidate its internal mechanisms. We compared the acoustic model of the BLSTM-HMM hybrid model with the encoder of the end-to-end model, and visualized them using t-SNE to examine the difference between neural network layers. As a result, we were able to delineate the difference between the acoustic model and the end-to-end model encoder. Additionally, we analyzed the decoder of the end-to-end model from a language model perspective. Finally, we found that improving end-to-end model decoder is necessary to yield higher performance.

Keywords: speech recognition, end-to-end, t-SNE, sequence-to-sequence

1. 서론

음성인식(automatic speech recognition)이란 입력으로 사람의 음성을 받아 어떤 대사를 말하였는지 인식하여 텍스트로 변환하고 이를 출력하는 것이다. 음성인식은 스마트폰에서부터 에어컨, 냉장고 그리고 AI 스피커에서 찾아볼 수 있다. 이렇듯 음성인식 기술은 단순히 음성 명령으로 기기를 조종하는 것에서 나아가 인공지능이 사람과 음성 대화를 나누기 위하여 필수적인 모듈로써 그 중요성이 높아지고 있다. 이러한 음성인식이 이루어지는 과정은 다음과 같다. 먼저 기기가 사람의 음성을 입력으로 받고, 기기 내부에서 이미 훈련되어 있는 음성인식 모델을 사용하여 입력 음성을 인식하고 텍스트로 변환한다. 이렇게 변환된 텍스트를 최종 출력으로 가지게 된다. 이때 사용하는 음성인식 모델은 다양한 구조를 가질 수 있지만, 가장 널리 사용되는 구조는 다음과 같다. 입력 음성이 특징 추출(feature extraction), 음향 모델(acoustic model) 그리고 언어 모델(language model)을 거쳐 최종적으로 단어열로 인식하게 된다. 특징 추출에서는 입력 음성을 짧은 시간 단위인 프레임(frame)으로 자르고 매 프레임마다 mel-frequency cepstral coefficients(MFCC), 필터 뱅크(filter bank) 등의 방법을 이용하여 수십 차원의 특징 벡터를 추출하여 음향 모델에 넘겨준다. 음향 모델은 입력으로 받은 특징 벡터를 음소(phoneme)로 인식하는 역할을 한다. 고전적으로 음향 모델은 가우시안 혼합 모델 기반 은식 마르코프 모델(Gaussian mixture model based hidden Markov model, GMM-HMM) 등을 사용한다.

* hoirkim@kaist.ac.kr, Corresponding author

Received 1 February 2019; Revised 22 February 2019; Accepted 28 February 2019

음향 모델이 입력 음성으로부터 음소열을 인식한 후에는, 언어 모델이 음소열을 단어열로 변환하는 역할을 수행한다. 이러한 언어 모델은 N-gram 등을 사용한다.

최근 심층신경망(deep neural network, DNN) 알고리즘이 다양한 머신러닝 분야에 사용되며 성능 향상이 이루어졌다. 음성 인식 분야에서도 역시 신경망을 사용하여 성능 향상이 크게 이루어졌음을 Hinton et al.(2012)의 연구에서 알 수 있으며, 이로 인해 최근에는 전통적인 음성인식 모델에 신경망을 적용한 하이브리드 모델이 주로 연구된다. 음향 모델의 경우 기존의 GMM-HMM 모델에서 GMM의 역할을 DNN으로 대체한 DNN-HMM 하이브리드 모델이 더 좋은 성능을 보여주는 것을 Hinton et al.(2012)의 연구에서 확인할 수 있다. 이후 DNN을 개선하여 convolutional neural network(CNN)나 recurrent neural network(RNN)를 사용한 모델들이 제안되었다(Abdel-Hamid et al., 2012; Graves et al., 2013). 언어 모델의 경우 Mikolov et al.(2010)의 연구에서 RNN 구조를 가진 RNNLM이 제안되었고 이는 많은 음성인식 연구에 적용되고 있다.

이와 같이 많은 연구를 통해 개선이 이루어지며 음성 인식의 성능이 향상되었지만, 최근에는 이러한 성능 개선의 방향이 점점 한계를 보이는 추세이다. 또한 음향 모델, 언어 모델 그리고 디코딩 서치를 거치는 일련의 음성 인식 과정에서 불필요한 정보가 개입되거나 필요한 정보가 제거될 수 있다는 문제가 존재한다. 더불어, 이미지 인식 또는 기계 번역과 같은 머신러닝을 사용하는 분야에서는 다양한 신경망을 조합하여 통합 신경망 모델을 사용하여 뛰어난 성능을 보이고 있다. 이러한 점들로 인해 음성 인식에서도 기존과 다른 새로운 음성 인식 모델의 필요성이 대두되었고, 그 결과로 종단간 음성인식 모델이 연구되기 시작하였다.

1.1. 종단간 음성 인식

종단간(end-to-end) 음성인식이란 입력 음성으로부터 하나의 통합 신경망을 거쳐 문자열 혹은 단어열을 인식하는 음성인식 방법을 뜻한다. 앞서 설명한 전통적인 음성 인식에서는 훈련을 위해 별도로 발음 사전을 준비해야 하고, 프레임의 정답 음소를 만들기 위한 여러 단계를 거쳐 음향 모델을 훈련시킨다. 또한 음향 모델과 발음 사전 그리고 별도로 훈련된 언어 모델을 결합하기 위해 weighted finite-state transducers(WFST)를 사용하여 가장 확률이 높은 경로를 찾아 최종적으로 음성을 인식하게 된다. 이러한 과정은 번거로울 뿐 아니라 각각의 역할을 이해하기 위해 음성 인식에 대한 사전 지식을 필요로 한다. 최근 심층신경망 기술의 발전에 따라 이와 같은 단점을 해결한 간단한 구조의 종단간 음성인식 모델이 제안되었다.

대표적인 종단간 음성인식 모델은 connectionist temporal classification(CTC) 방식이다(Graves et al., 2006). 이 방법은 재귀신경망을 이용해 음성 특징으로부터 바로 문자열을 추론해낸다. HMM과 유사하게 매 프레임마다 문자 사후확률을 추정하고, 이렇게 추정된 문자열이 최적의 경로를 갖도록 작동한다.

CTC와 다른 접근 방법으로는 기계번역 분야에서 주목할 만한 성능 향상을 이룬 sequence-to-sequence(seq2seq) 모델을 기반으로 한 음성인식 모델이 제안되었다(Chorowski et al., 2014; Miao et al., 2015; Sutskever et al., 2014). 이 모델은 재귀신경망으로 된 인코더(encoder)와 디코더(decoder)로 이루어져 있고, 입력 음성 특징으로부터 인코더가 매 프레임에 대하여 출력을 계산한다. 디코더는 어떤 프레임에 대한 인코더 출력에 주목(attention)할지 계산하여 주목도에 따라 인코더 값을 입력으로 사용하여 최종 문자열을 추정해낸다. 이 seq2seq 기반 음성인식 모델은 다른 종단간 음성인식 모델과 비교할 만한 성능을 보이며 지속적인 연구가 이루어지고 있다.

종단간 음성인식 모델의 장점으로는 그 구조가 이해하기 쉽다는 점이 있다. 전통적인 음성인식 구조와 다르게 통합된 신경망으로 이루어져 있고, 중간에 음소 인식을 거친 후 다시 텍스트로 변환하는 과정이 없이 입력을 바로 텍스트로 변환한다. 또한 훈련 과정이 간단하다는 점이 있다. 전통적인 음성 인식에서는 정답 텍스트뿐 아니라 훈련 및 인식에 사용하는 모든 단어에 대한 발음 사전과 별도로 훈련된 언어모델이 필요했다. 그리고 음향 모델 훈련을 위해 먼저 모노폰(monophone) 단위의 인식 모델을 만들고, 이후 이러한 모델을 통해 만든 음소 정렬을 정답으로 사용하여 트라이폰(triphone) 모델을 훈련하는 과정이 단계적으로 이루어져야 했다. 하지만 종단간 음성인식 모델은 입력 음성과 그에 대응되는 정답 텍스트만 가지고 훈련을 할 수 있다. 이렇듯 종단간 모델은 구조가 이해하기 쉽고 훈련하는데 별도의 정답을 만드는 등의 추가 과정이 필요 없다는 장점을 가진다.

종단간 음성인식 모델의 단점으로는 전통적인 구조의 하이브리드 음성인식 모델에 비하여 낮은 성능을 보이는 점이다. 또한 Dayhoff & DeLeo(2001)의 연구에서 신경망이 내부적으로 어떤 과정을 거쳐 인식하는지 이해하기 어렵다는 점을 언급하였다. 낮은 성능의 문제는 주로 언어 모델의 차이로 발생하게 된다. 언어 모델의 훈련에는 대부분 음성 데이터와 다른 훨씬 많은 분량의 텍스트를 사용하는데, 이에 비하여 음성 데이터와 음성 데이터의 정답 텍스트만을 사용하여 훈련한 종단간 음성인식 모델은 적은 데이터를 사용하는 셈이다. 따라서 이러한 단점을 보완하기 위하여 별도로 훈련한 언어 모델을 사용하여 성능을 개선하는 방법이 연구되고 있다(Chan et al., 2016; Miao et al., 2015; Sutskever et al., 2014). 또한 전통적인 음성인식에 사용되던 언어 모델을 종단간 음성인식 모델에 맞추어 다르게 적용하는 방법이 Sriram et al.(2017)의 연구에서 제안되었다. 내부적으로 인식이 어떻게 이루어지는지 알기 어려운 문제는 통합신경망을 사용하기 때문에 발생한다. 이는 비단 음성인식에서 뿐 아니라, 신경망을 사용하는 많은 머신러닝 분야에서 최근 중요하게 다뤄지고 있는 문제이다. 모델의 작동 방식을 알지 못한다는 것은 그 모델에 대한 신뢰도와 관련된 문제이기 때문에, 설명 가능한 인공지능(explainable AI)에 대한 연구가 이루어져야 함을 Gunning(2017)은 주장하였다. 본 논문에서는 이러한 관점에서 종단간 음성인식 모델의 내부적인 작동 원리를 파악하여 문제를 해결하고, 나아가 종단간 음성인식 모델의 성능 향상을 위한 방향을 제시하기 위하여 분석을 진행하였다.

1.2. 연구 목적

본 논문에서는 종단간 음성인식 모델의 시각화 분석을 진행하고, 분석 결과를 활용하여 성능 개선을 위한 방향을 제시한다. 앞서 기술한 바와 같이 종단간 음성인식 모델은 내부적으로 어떤 방식으로 음성 인식이 이루어지는지 알기 어렵다. 따라서 모델의 성능 개선을 위해서는 구조 변경이나 파라미터 튜닝을 통해 단어오류율(word error rate, WER) 등의 최종 성능만을 참고하게 된다. 이러한 방법의 성능 향상은 많은 시행착오를 거쳐야 하며, 모델이 다른 데이터베이스 환경에서 어떻게 작동할지 예측하기 어렵다. 이는 결과적으로 모델에 대하여 충분히 신뢰성을 가지기 어렵게 만든다. 그리고 일반적으로 종단간 음성인식은 전통적인 방식의 하이브리드 음성인식 모델에 비해 성능이 떨어지고, 이러한 단점을 보완하기 위해 기존의 음성인식 모델에서 사용되던 언어 모델을 사용하지만, 그럼에도 몇 가지 경우를 제외하고 하이브리드 음성인식 모델의 성능을 넘기기 어렵다.

따라서 이러한 문제점들을 해결하기 위하여 종단간 음성인식 모델이 어떠한 방식으로 음성인식을 행하는지 그 방식을 이해하고, 이를 활용하여 효과적인 성능 개선을 위한 방향 제시가 필요하다. 본 논문에서는 종단간 음성인식 모델이 작동 방식을 이해하기 위해 시각화 분석을 진행할 것이고, 분석 결과를 전통적인 하이브리드 모델과 비교하며 종단간 음성인식 모델의 특징을 바탕으로 성능 향상을 위한 방향을 제시할 것이다.

2. 음성 인식 모델의 분석 방법

2.1. 사용 DB

실험에 사용한 데이터는 Panayotov et al.(2015)이 발표한 Librispeech이다. LibriVox 영어 오디오북의 음성 데이터로 이루어졌으며, 음성 분량은 train-clean 460시간과 train-other 500시간, test-clean 5시간과 test-other 5시간의 음성 분량을 가지고 있다. 본 연구에서는 음성 인식 모델 훈련에 train-clean과 train-other의 960시간 분량을 모두 사용하였고, 시각화 분석에는 test-clean의 음성을 인식한 결과를 사용하였다. Librispeech는 16 kHz의 샘플링 레이트와 16 bit 심도를 가진 flac 무손실 압축 음성으로 이루어져 있다. 제공되는 언어 모델은 Project Gutenberg books에서 총 14,500권의 텍스트 분량을 사용하여 훈련되었으며, 3-gram에서 170의 perplexity와 0.4% 정도의 out of vocabulary를 가진다. 본 연구에서 하이브리드 음성 인식 모델의 성능 측정은 제공된 3-gram 언어모델을 기준으로 진행하였다.

2.2. 사용 툴킷

본 연구에서 하이브리드 음성 인식 모델을 구성하는데 사용한 툴킷(toolkit)은 Kaldi이며, 종단간 음성 인식 모델을 구성하는데 사용한 툴킷은 ESPNet이다. 본 연구는 Nvidia GTX1080Ti 4개와 Xeon E5 2620 v4를 가진 Ubuntu 16.04 환경의 서버에서 진행하였다.

ESPnet은 종단간 음성 처리를 위한 Python 기반 오픈 소스 플랫폼으로 2018년 공개되었다(Watanabe, 2018). 하이브리드 CTC-attention 기반 종단간 음성 인식과 Tacotron2 기반의 종단간 음성 합성을 Chainer 및 Pytorch를 이용하여 구현할 수 있다. 본 논문에서는 각 툴킷에서 제공하는 Librispeech에 대한 recipe를 참고하여 attention 기반 종단간 음성 인식 모델을 구현하였다.

2.3. 음성 인식 모델의 구조

분석에 사용한 음성 인식 모델은 세 가지로, BLSTM-HMM 음향 모델을 사용한 하이브리드 모델, 8층의 인코더와 2층의 디코더 구조인 종단간 모델 A 그리고 4층의 인코더와 6층의 디코더를 가진 종단간 모델 B이다. 하이브리드 모델은 전통적인 음성 인식 구조를 가지고 있으며 종단간 모델 A와 B와의 비교를 통해 그 차이를 알아보고자 하였다. 하이브리드 모델은 Kaldi, 종단간 모델은 ESPnet 툴킷을 이용하여 구현하였다.

BLSTM-HMM 하이브리드 모델은 Bi-directional LSTM 3층을 사용하여 음향 모델을 구성하였는데, 이때 각 BLSTM 레이어는 512차의 비재귀적 투영(projection) 벡터와 512차의 재귀적 투영 벡터, 합쳐서 총 1,024차의 투영 벡터를 출력으로 가진다. 그리고 마지막 레이어는 71개의 음소를 이용하여 만든 5,992개의 트라이폰 HMM state에 대한 출력 노드를 가진다. 기타 조건은 Kaldi 툴킷에서 제공하는 Librispeech recipe 중 chain model을 구현하였다. Librispeech에서 제공하는 3-gram 언어모델을 사용하여 테스트 클린 음성에 대하여 인식 성능을 측정한 결과, 단어오류율 4.23%와 문장오류율 43.4%를 얻을 수 있었다.

종단간 음성인식 모델은 location-aware attention 메커니즘을 사용하는 seq2seq 모델을 바탕으로 구현하였고, 그 구조는 그림 1과 같다.

종단간 모델 A는 인코더로 8층의 Bi-directional LSTM을 사용하였고 각 레이어는 320차의 투영 벡터를 출력으로 한다. Attention 메커니즘은 location-aware 알고리즘을 사용하였으며 window 크기는 5 프레임, 채널 크기는 10, 합성곱 필터 크기는 100 그리고 320차의 투영 벡터로 구성하였다. 디코더로 2층의 uni-directional LSTM을 사용하였고 각 레이어는 320차의 투영 벡터를 출력으로 한다. 최종 레이어에서는 30개의 문자 토큰(A-Z, ‘, blank, unk, eos, _)을 인식하도록 설정하였다.

종단간 모델 B는 인코더가 4층, 디코더가 6층으로 나머지 구조 및 매개변수는 종단간 모델 A와 같다.

종단간 모델에서 훈련 및 인식에 사용하는 문자 토큰은 다음과 같은 규칙으로 변형된다. 알기 쉽게 예를 들어 “I have a pen.”이라는 훈련용 문장을 “_I_HAVE_A_PEN<eos>”의 토큰 형태로 바꾸는 경우를 살펴본다. 먼저 문장의 처음에 문장의 시작을 뜻하는 토큰으로 “_”를 삽입한다. 이어서 알파벳은 그에 대응되는 대문자 알파벳 토큰으로, 공백은 “_” 토큰으로 대체한다. 그리고 아포스트로피를 제외한 기호는 삭제하고, 마지막으로 문장의 마지막에는 문장의 끝(end of sentence)을 뜻하는 “<eos>” 토큰을 삽입한다. 이렇게 생성된 토큰 형태의 문장을 정답으로 하여 종단간 음성 인식 모델의 훈련에 사용하고, 인식 결과로 생성된 토큰 형태의 문장을 규칙에 따라 역변환하여 타겟 문장과 비교한다.

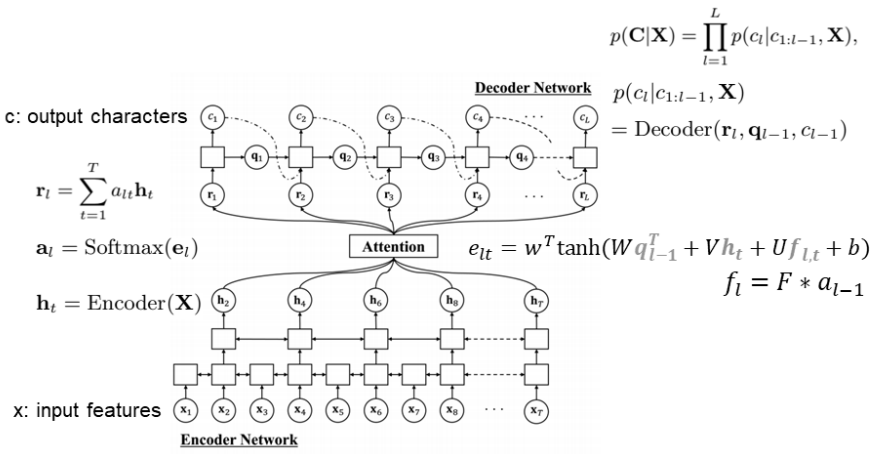

그림 1. Seq2seq 종단간 음성인식 모델의 구조
Figure 1. Structure of seq2seq end-to-end speech recognition model

2.4. 시각화 분석을 위한 차원축소 방법

고차원 데이터 세트는 시각화하여 분석하기 매우 어렵다. 따라서 데이터의 고유한 구조를 보여주기 위해 2차원 또는 3차원 그래프에 나타내는 것이 직관적으로 알아보기 쉽다. 이렇게 고차원 데이터 세트를 차원 축소하는 방법으로 몇 가지 방법이 연구되었다. 본 논문에서는 그 중 데이터의 군집 정도에 대한 분석에 특화된 차원 축소 방법인 t-distributed stochastic neighbor embedding(t-SNE)을 사용하였다.

t-SNE는 데이터의 유사도를 확률로 변환한다. 원래 공간에서의 유사도는 가우시안 분포로 표현되며, 저차원 공간에서의 유사도는 Student's *t* 분포로 표현된다. Student's *t* 분포는 가우시안 분포와 비슷한 모양을 가지고 있으며, 0을 기준으로 대칭을 이룬다. 하지만 가우시안 분포에 비해 양쪽 꼬리가 더 두터운 형태이다. t-SNE는 지역적 구조에 민감하여 데이터의 클러스터를 추출하는 경향을 가지고 있다. 이러한 지역적 특성을 기반으로 가깝게 위치한 데이터들을 뭉쳐 보이도록 표현하는 기능은 특히 숫자 데이터 집합이나 여러 고차원 구조를 동시에 포함하는 데이터 집합을 시각화하는데 효과적이다. t-SNE에서는 고차원 공간과 저차원 공간에서의 유사도를 Kullback-Leibler(KL) 발산을 통해 최소화한다. 이 때문에 초기화를 할 때마다 다른 결과를 보여줄 수 있다. 또한 t-SNE는 다른 차원 축소 방법에 비해 계산 비용이 많이 들고, 전체적인 수학적 구조는 보존되지 않는다. 따라서 전체적인 구조를 어느 정도 보존하기 위해 principal component analysis(PCA)를 먼저 어느 정도 수행하고 진행하는 경우도 있다. 그림 2는 t-SNE를 사용하여 필기체 숫자 집합인 MNIST(LeCun, 1998)의 데이터를 시각화한 모습의 예이다.

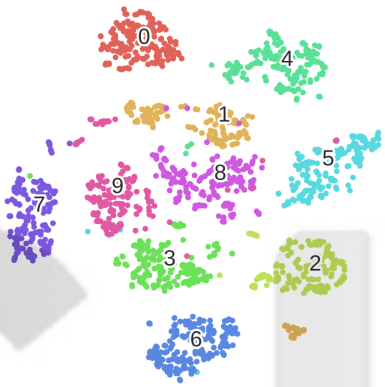

그림 2. t-SNE를 이용한 MNIST 시각화
Figure 2. Visualization of MNIST using t-SNE t-SNE, t-distributed stochastic neighbor embedding

3. 분석 결과

t-SNE 시각화 분석에서 두 가지 방법으로 차원 축소된 벡터에 대하여 색상별로 표시하였다. 첫 번째 방법은 시간 순으로 프레임을 나타낸다. 시간 순으로 무지개 색에 대응하여 음성의 처음 프레임은 붉은 색, 음성의 마지막 프레임 색은 보라색으로 대응하여 표시하였다. 그리고 그래프 오른쪽의 색상 막대에는 프레임에 대응되는 문자열의 순서를 표기하였다. 이를 시간순서 라벨이라고 나타낸다.

두 번째로 특정 프레임이 어떤 음소에 해당하는지 알기 위하여 BLSTM-HMM 하이브리드 모델을 사용하여 강제 정렬(forced alignment)을 수행하였다. 이를 통해 음성의 특정 프레임이 어떠한 음소에 해당하는지 음소 정렬 정보를 얻을 수 있다. 이러한 정렬 정보를 활용하여 각 프레임이 어떠한 음소에 대응되는지 색상별로 표기하였다. 이러한 표기 방식을 음소 라벨이라고 나타낸다. 이때 사용한 발음사전은 Librispeech에서 제공되는 것을 사용하였으며 각 색상에 대응되는 음소는 발음사전의 모든 음소 중에서 분석한 특정 음성의 문장에서 사용된 음소만을 표기하였다. 이때 음소 중 SIL은 묵음 구간(silence)을 뜻한다.

3.1. 종단간 모델의 구조 변화에 따른 성능 비교

표 1. 종단간 모델의 음성 인식 성능 비교
Table 1. Comparison of end-to-end model performance

	End-to-End model A	End-to-End model B
Encoder layer	8	4
Decoder layer	2	6
WER (SER), %	10.3(61.1)	16.1(76.0)

성능은 표 1과 같다. 종단간 모델 A는 BLSTM 인코더 8층에 LSTM 디코더 2층을 사용하는 반면 종단간 모델 B는 BLSTM 인코더 4층에 LSTM 디코더 6층을 사용한다는 차이가 있다. 그 결과 인코더의 레이어 수가 감소하고 디코더 레이어 수가 증가한 종단간 모델 B에서 인식 성능이 저하되었음을 알 수 있다.

3.2. BLSTM-HMM 음향 모델과 종단간 모델 인코더의 시각화 분석

BLSTM-HMM 음향 모델과 종단간 모델 인코더의 시각화를 진행하기 위해 test-clean 데이터에서 특정 화자의 음성 모두를 활용하여 t-SNE를 훈련시키고 그 중 특정 문장만을 시각화하도록 했다. 그림 3에서부터 그림 6까지의 인식 문장은 동일 화자의 “oh that made him so angry”에 대하여 복문으로 t-SNE를 훈련시키고 분석한 결과이다.

그림 3과 그림 4를 보면 비슷한 색상을 가진 점들이 모여 있는 것을 확인할 수 있다. 이때 색상은 시간에 따라 대응하므로, 비슷한 색상의 점들은 비슷한 시간에 위치한 프레임들을 의미한다. 이때 모여서 클러스터를 이루는 모습은 최종 레이어에서 모인 프레임들이 비슷한 출력을 가짐을 의미한다.

그림 3과 그림 5에서 같은 위치의 점은 동일한 프레임을 나타내며, 그림 4와 그림 6 역시 그러하다. 이때 클러스터를 이루는 프레임들을 음소 라벨에서 살펴보면 수 개의 음소 구간임을 알 수 있다. 즉 음향 모델과 인코더가 수 개의 음소 구간을 묶어 하나의 의미로써 인식함을 알 수 있다.

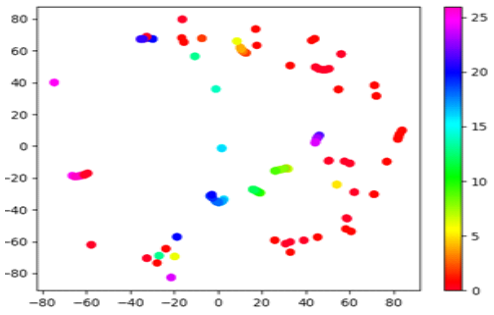

그림 3. BLSTM-HMM layer3 시간순서 라벨
Figure 3. BLSTM-HMM layer3 order label

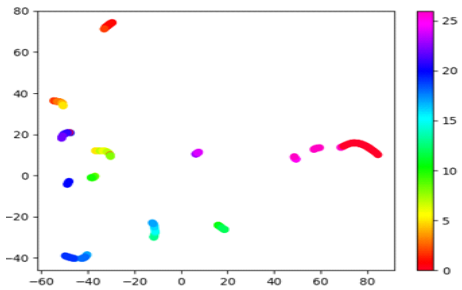

그림 4. 종단간 모델 A 인코더 layer 8 시간순서 라벨
Figure 4. Encoder layer8 of E2E model A, order label

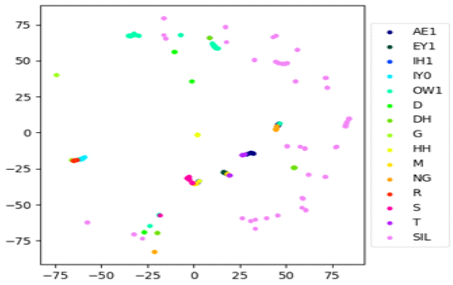

그림 5. BLSTM-HMM layer3 음소 라벨
Figure 5. BLSTM-HMM layer3 phone label

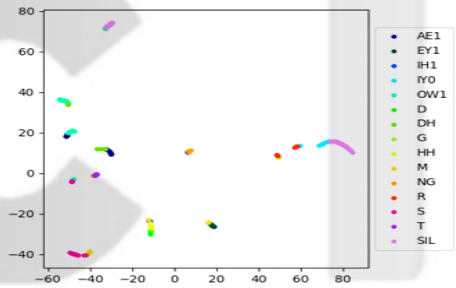

그림 6. 종단간 모델 A 인코더 layer 8 음소 라벨
Figure 6. Encoder layer8 of E2E model A, phone label

하이브리드 모델과 종단간 모델의 차이점은 그림 5와 그림 6을 통해 알 수 있다. 그림 5에서 하이브리드 모델의 경우 SIL음소, 즉 묵음 구간의 프레임이 넓게 흩어지는 경향이 있으며 나머지 음소 구간도 일부 흩어져 있음을 확인할 수 있다. 이와 다르게, 그림 6의 종단간 모델에서는 묵음 구간을 포함한 대부분의 프레임이 클러스터를 이루는 모습을 보인다. 두 모델은 공통적으로 BLSTM 구조를 사용함에도 이런 차이가 발생하였다. 따라서 종단간 모델이 하이브리드 모델과 비교하여 더욱 넓은 전후 프레임 관계를 모델링하며, 묵음 구간을 포함한 전후 발음에 대한 문맥적 언어 정보를 인식에 활용한다는 점을 알 수 있다. 이러한 차이는 음향 모델과 다르게 음소 단위의 정답을 거치지 않는 종단간 모델의 특성 때문으로 생각된다.

따라서 종단간 모델의 성능 개선을 위해서는 인코더를 전통적인 음성 인식에서의 음향 모델과 다르게 전후 문맥 정보를 모델링한다는 점을 고려하여야 한다. 따라서 인코더 개선을 위해

신경망 구조를 변경한다면, 전후 프레임 정보를 넓게 활용할 수 있는 재귀적 신경망이나 합성곱 신경망과 같은 구조를 사용하는 방향이 성능 개선에 도움이 될 수 있을 것이다. 또한 인코더의 문맥 모델링 특성을 효율적으로 활용한다면 성능 개선에 도움이 될 수 있을 것이다.

3.3. BLSTM-HMM 음향 모델과 종단간 모델 인코더의 레이어별 분석

신경망 모델에서 음성 인식에 사용하는 것은 최종 레이어의 출력이지만, 내부적으로 어떻게 음성인식을 수행하는지 이해하기 위해서는 하이브리드 모델과 종단간 모델의 BLSTM 신경망에서 레이어 별로 어떤 과정을 거쳐 인식이 이루어지는지 살펴봐야 한다. 이번 절에서는 수 개의 레이어 출력을 비교하여 각 레이어에서 어떤 정보를 활용하는지 알아보았다.

그림 7과 그림 8은 “at last the little mice stayed away also and the tree sighed after all it was very pleasant when the sleek little mice sat round me and listened to what I told them”이라는 음성을 t-SNE를 사용하여 분석한 결과이다. 이때 t-SNE는 해당 음성의 화자가 발성한 모든 음성을 사용하여 훈련하였다.

그림 7은 하이브리드 모델의 BLSTM 레이어 1, 2, 3층의 출력을 음소 라벨로 나타낸 것이다. 그림 8은 종단간 모델 A의 BLSTM 레이어 1, 3, 8층의 출력을 음소 라벨로 나타낸 것이다. 하이브리드 음향 모델의 1층 레이어의 그래프에서 AO1, AW, T, SIL등의 음소를 보면 비슷한 음소끼리 어느 정도 뭉쳐 있는 것을 확인할 수 있다. 하지만 2, 3층으로 올라가면서 이러한 음소끼리 뭉치는 경향이 감소함을 알 수 있다. 이러한 레이어별 차이는 종단간 모델에서 더욱 뚜렷하게 나타난다. 1층의 음소 라벨을 보면 비슷한 색, 즉 같은 음소끼리 뭉쳐 있는 경향을 확인할 수 있으며 SIL 역시 뭉쳐 있는 것을 확인할 수 있다. 이렇게 비슷한 음소끼리 뭉치는 모습은 3층 레이어에서도 나타난다. 이를 통해 종단간 모델의 인코더가 저층 레이어부터 중간층 레이어까지 음소 단위로 모델링을 한다는 것을 확인할 수 있다. 하지만 8층 레이어의 그래프를 보면 음소 단위와 관계없이 뭉치는 것을 알 수 있다. 즉 종단간 모델의 인코더가 낮은 레이어에서 중간 레이어까지는 음소 단위로 모델링하고, 이후 3.2.절에서 확인한 결과와 같이 최종 레이어에서는 전후 프레임과의 상관관계를 고려하여 수 개의 전후 음소가 묶인 음소열 단위로 음성 인식을 해낸다는 것으로 이해할 수 있다.

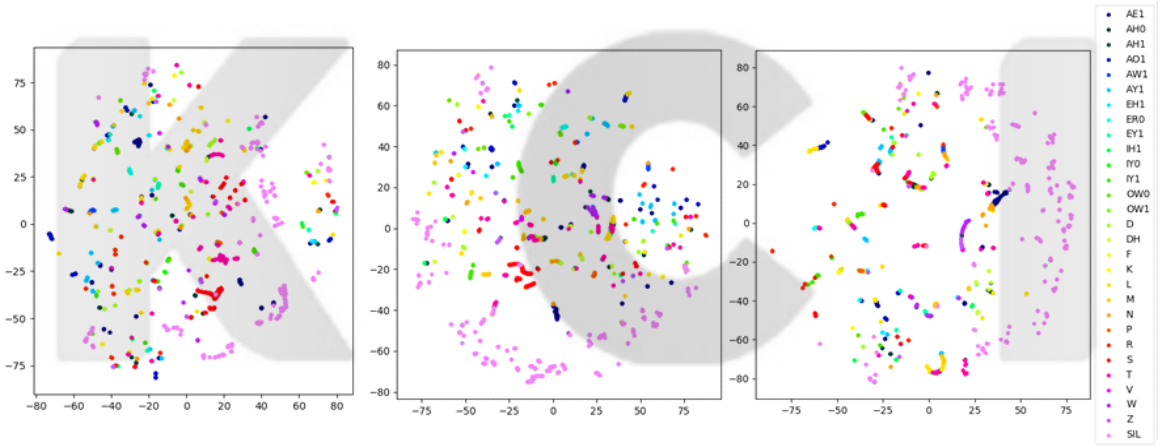

그림 7. BLSTM-HMM layer1(왼쪽), 2, 3(오른쪽) 음소 라벨
Figure 7. BLSTM-HMM layer 1(left), 2, 3(right) phone label

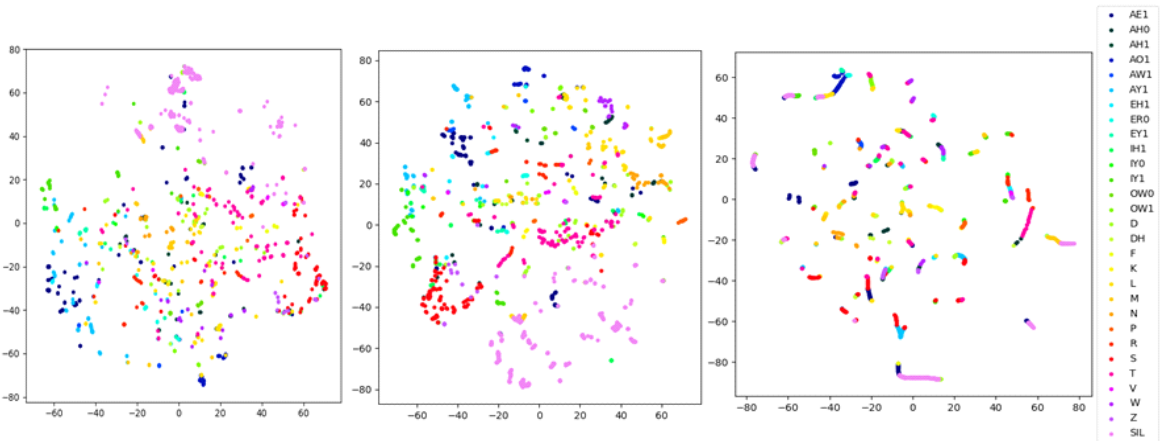

그림 8. 종단간 모델 A 인코더 layer 1(왼쪽), 3, 8(오른쪽) 음소 라벨
Figure 8. Encoder layer 1(left), 2, 3(right) of E2E model A, phone label

3.4. 종단간 모델 디코더에 대한 최근접 이웃 분석 및 비교

그림 9와 그림 10은 종단간 모델 B를 이용하여 인식 문장에 대한 각 문자의 디코더 출력의 최근접 이웃 5-best를 나타낸 것이다. 두 그림에서 인식 문장의 '<<', '>>' 두 기호 사이에 위치하는 문자가 기준 문자이며, 최근접 이웃의 '<<', '>>' 두 기호 사이에 위치하는 문자가 기준 문자의 최근접 이웃이다. 그림 9의 6층 레이어에서 좌측 상단 S 기준 문자를 비교한 경우를 보면, 최근접 이웃 문자가 모두 'YES'라는 단어의 마지막 S임을 확인할 수 있다. LSTM의 특성상 현재 기호보다 이전에 존재하는 문자를 이용하여 모델링하게 되는데, 이러한 모델링이 잘 이루어졌다고 할 수 있다. 다음의 '_' 토큰 역시 'YES' 단어를 모델링하고 있음을 확인할 수 있다. 하지만 'THEY SAID'에서의 'S' 경우에는 앞의 'THEY' 단어까지만 모델링하고 그 전의 'SO'는 최근접 이웃에 나타나지 않는다. 또한 바로 다음의 'A' 경우를 보면, 앞의 'S'뿐 아니라 앞의 'THEY'가 동일하게 등장하는 2개의 최근접 이웃을 확인할 수 있다. 하지만 다른 최근접 이웃은 'THEY'와 전혀 다른 단어를 앞에 가지고 있는 것을 알 수 있다. 그림 10의 디코더 1층에서는 최근접 이웃 5-best중 어떤 후보도 'THEY'를 모델링하고 있지 않다. 즉 단어 단위로는 이전의 최대 1개 단어에 대한 정보를 모델링하고, 또는 문자가 포함된 단어만 모델링한다는 점을 알 수 있다.

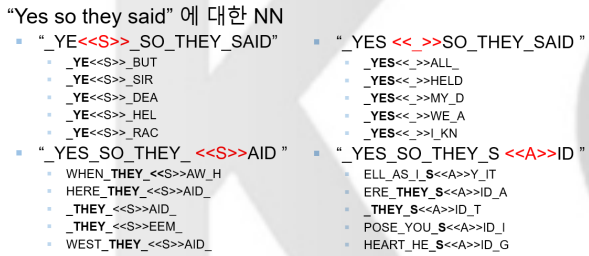

그림 9. 종단간 모델 B 디코더 layer6의 최근접이웃
Figure 9. Nearest Neighbors of Decoder layer 6 in E2E model B

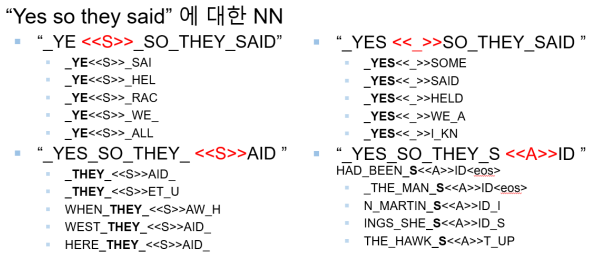

그림 10. 종단간 모델 B 디코더 layer1의 최근접이웃
Figure 10. Nearest Neighbors of Decoder layer 1 in E2E model B

3.5. 종단간 모델 디코더의 레이어별 기능 분석

디코더의 레이어별 차이를 정량적으로 분석하기 위하여 최근접 이웃 5-best 문장에서 기준 문자의 앞 문자가 원본 문자열과 몇 개까지 동일한지 통계적으로 조사해 보았다. 종단간 모델 B를 이용하여 총 79개 문장 9,566개 문자 토큰에 대하여 조사하였고, 그 결과는 표 2와 같다.

표 2를 보면 종단간 모델 디코더의 레이어별 기능 차이가 크게 나타나지 않음을 알 수 있다. 즉 이러한 경우 디코더의 레이어 증가는 성능에 도움이 크게 되지 않는다고 생각할 수 있다. 또한 평균적으로 앞 2개 문자에 대해서 언어 모델링이 이루어지는 것을 확인할 수 있다. 이는 기존의 언어 모델 역할을 수행하기에는 상당히 부족하다. 이러한 현상은 디코더가 언어 모델 역할을 수행하기에는 훈련에 사용한 데이터의 수가 너무 적기 때문에 나타났을 수 있다. 전통적인 음성인식 모델에서 사용하는 언어 모델의 경우 음성 데이터보다 훨씬 많은 양의 텍스트 데이터로 훈련하는 반면, 종단간 모델의 디코더는 인코더와 통합하여 음성 데이터로 훈련하게 된다. 따라서 전통적인 음성 인식에서 사용하는 언어 모델의 역할을 하기에는 훈련 데이터가 부족하여 적은 수의 레이어만 사용하는 것으로도 충분한 것이다. 또는 LSTM 신경망의 구조가 모델링에 적합하지 않을 수 있다. 따라서 종단간 음성인식 모델의 성능 향상을 위해 디코더가 더 많은 텍스트 데이터를 학습하도록 구조를 변경하거나 언어모델을 추가적으로 사용하여 인코더의 정보를 효율적으로 사용하여야 할 것이다.

표 2. 종단간 모델 B 디코더의 레이어별 성능 비교
Table 2. Comparison of E2E model B decoder performance by layer

Layer	동일 토큰 개수	평균 동일 토큰 수
1	88,024	1.84
2	93,080	1.94
3	95,254	1.99
4	94,084	1.96
5	92,401	1.93
6	92,625	1.93

4. 결론

본 논문에서는 attention 기반 종단간 음성 인식 시스템의 개선을 위한 시각화 분석을 진행하였고, 이를 통해 종단간 모델이 전통적인 방식의 하이브리드 모델과 어떻게 다르게 작동하는지 이해하고, 성능 개선을 위한 방향을 제안하였다. Librispeech 음성 DB로 훈련한 BLSTM-HMM 하이브리드 음성인식 모델의 음향 모델과 attention 기반 종단간 음성인식 모델의 인코더를 분석하기 위하여 각각의 BLSTM 레이어 출력을 t-SNE로 차원 축소 및 시각화하였다. 그 결과 종단간 모델의 인코더는 앞뒤 문맥을 고려하는 경향이 더욱 강하며 묵음 구간에서 역시 그러한 경향이 있음을 확인하였다.

종단간 모델의 디코더는 언어 모델 관점에서 그 역할을 얼마나 수행하는지 알아보기 위해 인식된 문장의 디코더 각 레이어의 출력을 최근접 이웃 탐색을 통해 다른 문장의 디코더 출력과 비교해 보았다. 이를 통해 문자가 속한 단어와 최대 앞의 1개 단어까지 모델링이 이루어지는 점과 평균적으로 기준 문자의 앞 2개 문자에 대하여 모델링이 이루어진다는 점을 확인할 수 있었다.

결론적으로, 종단간 모델은 전통적인 음성 인식 모델과 다르게 작동함을 알 수 있었으며, 종단간 모델의 인코더가 앞뒤 문맥을 모델링하여 인식을 진행한다는 점을 확인할 수 있었다. 또한 종단간 모델의 디코더가 언어 모델링에서 부족함을 확인하고 이에 대한 성능 개선 방안으로 디코더의 구조 변경 또는 언어 모델링 향상을 위한 텍스트 데이터 훈련 방법이 필요함을 제시하였다.

References

Abdel-Hamid, O., Mohamed, A. R., Jiang, H., & Penn, G. (2012, March). Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. *Proceedings of the 2012 IEEE International Conference on Acoustics*, *Speech and Signal Processing (ICASSP)* (pp. 4277-4280).

Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016, March). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. *Proceedings of the 2016 IEEE International Conference on Acoustics*, *Speech and Signal Processing (ICASSP)* (pp. 4960-4964).

Chorowski, J., Bahdanau, D., Cho, K., & Bengio, Y. (2014). End-to-end continuous speech recognition using attention-based recurrent NN: First results [Computing Research Repository]. Retrieved from http://arxiv.org/abs/1412.1602

Dayhoff, J. E., & DeLeo, J. M. (2001). Artificial neural networks: Opening the black box. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, *91*(S8), 1615-1635.

Graves, A., Fernandez, S., Gomez, F., & Schmidhuber, J. (2006, June). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. *Proceedings of the 23rd International Conference on Machine Learning* (pp. 369-376).

Graves, A., Mohamed, A. R., & Hinton, G. (2013, May). Speech recognition with deep recurrent neural networks. *Proceedings of the 2013 IEEE International Conference on Acoustics*, *Speech and Signal Processing (ICASSP)* (pp. 6645-6649).

Gunning, D. (2017). Explainable artificial intelligence (XAI). *Defense Advanced Research Projects Agency (DARPA)*.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A. R., Jaitly, N., Senior, A., ... Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, *29*(6), 82-97.

LeCun, Y., Cortes, C., & Burges, C. J. C. (1998). The MNIST database of handwritten digits. Retrieved from http://yann. lecun. com/exdb /mnist

Miao, Y., Gowayyed, M., & Metze, F. (2015, October). EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding. *Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (pp. 167-174).

Mikolov, T., Karafiat, M., Burget, L., Cernocky, J., & Khudanpur, S. (2010). Recurrent neural network based language model. *Proceedings of the 11th Annual Conference of the International Speech Communication Association* (pp. 1045-1048).

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015, April). Librispeech: An ASR corpus based on public domain audio books. *Proceedings of the 2015 IEEE International Conference on Acoustics*, *Speech and Signal Processing (ICASSP)* (pp. 5206-5210).

Sriram, A., Jun, H., Satheesh, S., & Coates, A. (2017). Cold fusion: Training seq2seq models together with language models [Computing Research Repository]. Retrieved from http://arxiv.org/abs/1708.06426

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Proceedings of the Advances in Neural Information Processing Systems* (pp. 3104-3112).

Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Enrique Yalta Soplin, N., ... Ochiai, T. (2018). ESPnet: End-to-End Speech Processing Toolkit [Computing Research Repository]. Retrieved from http://arxiv.org/abs/1804.00015

• **임성민 (Seongmin Lim)**
한국과학기술원 전기및전자공학부 석사
대전광역시 유성구 대학로 291
Tel: 010-7710-4835
Email: smlim01@kaist.ac.kr
관심분야: 음성인식

• **구자현 (Jahyun Goo)**
한국과학기술원 전기및전자공학부 박사과정
대전광역시 유성구 대학로 291
Tel: 010-5314-0598
Email: jahyun.goo@kaist.ac.kr
관심분야: 음성인식

• **김회린 (Hoirin Kim)** 교신저자
한국과학기술원 전기및전자공학부 교수
대전광역시 유성구 대학로 291
Tel: 042-350-7417
Email: hoirkim@kaist.ac.kr
관심분야: 음성인식, 음성합성, 화자인식

어텐션 기반 엔드투엔드 음성인식 시각화 분석

임 성 민 · 구 자 현 · 김 회 린

한국과학기술원 전기및전자공학부

국문초록

전통적인 음성인식 모델은 주로 음향 모델과 언어 모델을 사용하여 구현된다. 이때 음향 모델을 학습시키기 위해서는 음성 데이터에 대한 정답 텍스트뿐만 아니라 음성인식에 사용되는 단어의 발음사전과 프레임 단위의 음소 정답 데이터가 필요하다. 이 때문에 모델을 훈련하기 위해서는 먼저 프레임 단위의 정답을 생성하는 등의 여러 과정이 필요하다. 그리고 음향 모델과 별도의 텍스트 데이터로 훈련한 언어 모델을 적용하여야 한다. 이러한 불편함을 해결하기 위하여 최근에는 하나의 통합 신경망 모델로 이루어진 종단간(end-to-end) 음성인식 모델이 연구되고 있다. 이 모델은 훈련에 여러 과정이 필요없고 모델의 구조를 이해하기 쉽다는 장점이 있다. 하지만 인식이 내부적으로 어떤 과정을 거쳐 이루어지는지 알기 어렵다는 문제가 있다. 본 논문에서는 어텐션 기반 종단간 모델을 시각화 분석하여 내부적인 작동 원리를 이해하고자 하였다. 이를 위하여 BLSTM-HMM 하이브리드 음성인식 모델의 음향 모델과 종단간 음성인식 모델의 인코더를 비교하고, 신경망 레이어 별로 어떠한 차이가 있는지 분석하기 위해 t-SNE를 사용하여 시각화하였다. 그 결과로 음향모델과 종단간 모델 인코더의 차이점을 알 수 있었다. 또한 종단간 음성인식 모델의 디코더의 역할을 언어모델 관점에서 분석하고, 종단간 모델 디코더의 개선이 성능 향상을 위해 필수적임을 알 수 있었다.

핵심어: 음성 인식, 엔드투엔드, 종단간, t−SNE, 시퀀스 투 시퀀스