pISSN 2005-8063
eISSN 2586-5854
2020. 6. 30.
Vol.12 No.2
pp. 29-37

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2020.12.2.029

Semi-supervised domain adaptation using unlabeled data for end-to-end speech recognition

Hyeonjae Jeong · Jahyun Goo · Hoirin Kim*

School of Electrical Engineering, KAIST, Daejeon, Korea

Abstract

Recently, the neural network-based deep learning algorithm has dramatically improved performance compared to the classical Gaussian mixture model based hidden Markov model (GMM-HMM) automatic speech recognition (ASR) system. In addition, researches on end-to-end (E2E) speech recognition systems integrating language modeling and decoding processes have been actively conducted to better utilize the advantages of deep learning techniques. In general, E2E ASR systems consist of multiple layers of encoder-decoder structure with attention. Therefore, E2E ASR systems require data with a large amount of speech-text paired data in order to achieve good performance. Obtaining speech-text paired data requires a lot of human labor and time, and is a high barrier to building E2E ASR system. Therefore, there are previous studies that improve the performance of E2E ASR system using relatively small amount of speech-text paired data, but most studies have been conducted by using only speech-only data or text-only data. In this study, we proposed a semi-supervised training method that enables E2E ASR system to perform well in corpus in different domains by using both speech or text only data. The proposed method works effectively by adapting to different domains, showing good performance in the target domain and not degrading much in the source domain.

Keywords: automatic speech recognition, end-to-end, semi-supervised, domain adaptation

1. 서론

음성인식(automatic speech recognition, ASR)이란 사람의 음성 신호를 입력으로 받아 그 발화 내용을 인식하여 문자열로 변환하고 이를 출력하는 것을 말한다. 음성인식 시스템은 다양한 구조를 가질 수 있지만, 고전적으로 널리 사용되는 음성인식 시스템의 구조는 입력 음성을 시스템에서 사용하기 용이하도록 변환해 주는 특징 추출 과정, 추출된 특징을 이용해 발화된 음소열을 판별해 주는 음향모델, 최종적으로 음소열을 문자열로 변환해 주는 언어모델로 이루어져 있다. 특징 추출 과정에서는 음성 신호를 짧은 시간 단위인 프레임(frame)으로 자르고 매 프레임마다 MFCC, mel-filterbank 등의 방법을 이용하여 프레임별 특징을 추출하게 되고 수십 차원의 특징 벡터를 음향모델 입력으로 넣어주게 된다. 음향모델에서는 특징 추출 과정에서 받은 특징 벡터를 음소열로 바꾸기 위해 통계적 모델을 사용하며, 고전적인 음성인식 시스템에서 사용하는 음향모델은 가우시안

* hoirkim@kaist.ac.kr, Corresponding author

Received 10 February 2020; Revised 4 May 2020; Accepted 8 May 2020

혼합 모델 기반 은닉 마르코프 모델(Gaussian mixture model based hidden Markov model, GMM-HMM)이 주로 사용된다. 음향모델에서 특징 벡터가 음소열로 변환된 이후에는 언어모델이 이를 문자열로 변환해 주는 역할을 수행한다. 언어모델은 N-gram 기반 시스템이 가장 널리 사용되며 문자열 출력을 단어열 단위로 출력해 주게 된다.

음성인식 시스템의 훈련은 음향모델과 언어모델을 훈련하는 것을 말하며 이는 다음과 같이 이루어진다. 먼저 입력 음성 신호와 그에 대한 정답 문자열을 준비한다. 또한 정답 문자열에 포함된 모든 단어들에 대해 각각 음소열로 대응시킨 발음사전(lexical dictionary)을 준비한다. 고전적인 음성인식 시스템에서 사용되는 GMM-HMM 음향모델을 훈련하기 위해서는 음성에서 추출된 특징 벡터에 각 프레임 별로 대응되는 정답 음소가 필요하다. 이를 얻기 위해서 비터비 정렬(Viterbi alignment)을 거쳐 비교적 간단한 모델인 모노폰(monophone) 음향모델을 훈련하고, 이 음향모델로 얻어지는 정답 음소열 데이터를 이용하여 조금 더 복잡한 모델인 트라이폰(triphone) 음향모델을 훈련하는 과정을 거친다. 이렇게 순차적으로 학습된 음향모델은 입력 특징 벡터로부터 그에 대응하는 음소열을 출력할 수 있게 된다. 다음으로 N-gram 언어모델의 훈련은, 사용하고자 하는 말뭉치에 있는 정답 문자열에 있는 단어를 포함하여 더 많은 양의 방대한 언어모델 훈련용 문자열 데이터를 활용하여 이루어진다. 언어모델은 전체 문자열 데이터에서 특정 단어가 등장할 확률과 앞 단어들을 고려한 조건부 확률을 계산하고 저장한다. 이를 통해 단어열로 이뤄진 문장이 등장할 빈도를 확률적으로 계산할 수 있다. 최종적으로 훈련된 음향모델과 언어모델을 결합하여 인식 단어열의 확률을 계산하는 과정을 디코딩 서치(decoding search)라고 하며, 이는 weighted finite state transducer (WFST)를 통해 이뤄지게 된다.

최근에는 심층학습 기반 알고리즘이 다양한 머신러닝 분야에 사용되며 성능이 비약적으로 향상되었다. 음성인식 분야에서도 심층신경망(deep neural network, DNN)을 이용한 음향모델이 고전적인 GMM-HMM 기반 음향모델보다 성능이 뛰어나다는 것이 확인되었고(Hinton et al., 2012), GMM의 역할을 DNN이 대체하는 음향모델인 DNN-HMM 음향모델을 사용하는 하이브리드 음성인식 연구들이 많이 제안되었다. 이후 컨볼루션 신경망(convolutional neural network, CNN)이나 재귀신경망(recurrent neural network, RNN) 등의 개선된 신경망 모델이 등장했으며 더욱 나은 성능을 보여주고 있다(Graves et al., 2013). 언어모델에서도 역시 RNN 기반으로 훈련된 모델이 좋은 성능을 보여 연구가 진행되고 있다(Mikolov et al., 2010). 하지만 최근에는 이러한 방향의 성능 개선이 점차 한계를 보이고 있다. 또한 음성인식 시스템에 필요한 각 모델을 따로 훈련하여 비터비 정렬, 디코딩 서치 등을 이용해야 한다는 단점이 부각되어 음성인식에서도 기존과 다른 새로운 음성인식 모델이 필요해졌고, 모든 모델을 통합하여 하나의 거대한 모델 하나만을 사용해 훈련할 수 있는 음성인식 시스템이 제안되었으며, 이것이 종단간 음성인식 시스템이다.

1.1. 종단간 음성인식

종단간(end-to-end) 음성인식이란 입력 음성을 입력받아 하나의 통합 신경망을 거쳐 문자열이나 단어열로 출력하는 음성인식 방식을 말한다. 앞서 설명한 고전적인 음성인식 시스템을 훈련하기 위해서는 별도로 준비된 발음사전이 필요하며 각 프레임별 정답 음소열을 구하기 위해 몇 차례의 음향모델 훈련이 요구되며 또한 음향모델과 별개로 훈련된 언어모델과 통합하기 위해 WFST 기반 디코딩 서치가 필요하다. 이러한 과정은 번거로울 뿐만 아니라 각 모델의 역할을 이해하기 위해 음성인식에 대한 구체적인 지식을 요구한다. 최근 심층학습 기법의 장점을 더욱 잘 활용하는 방법으로 언어모델링 및 디코딩 과정을 통합해 처리하는 종단간 음성인식 시스템이 제안되었다. 가장 먼저 제안된 종단간 음성인식 모델은 conntctionist temporal classification (CTC) 기법이다(Graves et al., 2006). 이 방법은 재귀신경망을 이용해 음성 특징 벡터로부터 문자열 출력을 직접 얻는다. 이는 HMM과 비슷하게 매 프레임마다 문자가 등장할 사후확률을 추정하고 추정된 문자열이 최적의 경로를 갖도록 작동한다. 이후 CTC 모델의 성능을 개선하기 위해 다양한 연구가 진행되었다. WFST를 이용해 기존 사용되던 언어모델과 결합하여 디코딩 하는 방법이나 재귀신경망 이외의 신경망 모델을 사용하는 방법을 통해 음성인식 성능의 향상이 보고되었다(Miao et al., 2015). 이후 신경망 기반 기계번역 분야에서 높은 성능 향상을 보인 어텐션 기반 시퀀스투시퀀스(attention-based sequence-to- sequence) 모델을 음성인식기에 적용한 방법이 제안되었다(Chan et al., 2016). 이 모델은 재귀신경망 기반의 인코더(encoder)와 디코더(decoder)로 이루어져 있다. 인코더는 입력 음성 특징으로부터 매 프레임에 대해 출력을 계산한다. 디코더는 모든 프레임에 대한 인코더의 출력을 통째로 전달받으며 현재 출력을 계산할 때 인코더의 어떤 프레임을 주목(attention)할지 학습하여 그에 따라 프레임별 인코더 값을 다른 비중으로 계산해 최종 문자열을 추정한다. 시퀀스투시퀀스 모델 또한 큰 성능향상을 보이며 지속적인 연구가 이루어지고 있다.

종단간 음성인식 시스템은 기존의 고전적인 음성인식 시스템이나 하이브리드 음성인식 시스템에 비해 몇 가지 장점을 갖고 있다. 먼저 그 구조가 간결하다는 것이 장점이다. 고전적인 음성인식 시스템은 여러 개의 모델로 구성되어 있으며, 각 모델을 통합하기 위해 복잡한 기법을 사용한다. 하지만 이와 달리 종단간 음성인식 시스템은 하나의 큰 통합 모델에 음성 신호를 입력하면 출력으로 문자열을 바로 얻을 수 있으며 모델 통합을 위한 추가적인 계산이 필요하지 않다. 또한 종단간 음성인식 시스템은 한 번의 훈련으로 전체 음성인식 시스템이 훈련된다는 것이 장점이다. 고전적인 음성인식 시스템은 음향모델과 언어모델을 각각 훈련해야 하며, 음향모델 훈련에는 언어모델과의 통합을 위해 프레임별 음소열을 출력하기 위해 몇 번의 단계적인 훈련이 필요했다. 하지만 종단간 음성인식 시스템은 입력 음성과 그에 대응하는 정답 문자열만으로 모든 훈련이 가능하며, 이는 음성인식에 대한 진입장벽을 낮추고 모델 통합에서 발생하는 정보 손실도 피할 수 있다. 또한 모델의 크기를 쉽게 조정

할 수 있어 비교적 저장용량이 적은 기기에도 손쉽게 적용할 수 있다는 장점이 있다.

하지만 종단간 음성인식기에도 단점은 존재한다. 하나의 큰 모델에서 모든 인식 과정이 일어나지만 어떤 방식으로 음성인식이 이루어지는지 알기 어려워 성능 개선 방향을 찾기 어렵다는 점과, 훈련 데이터의 양이 한정될 경우 기존 구조의 DNN-HMM 음성인식기에 비해 낮은 성능을 보이는 것이 단점이라고 할 수 있다. 이를 해결하기 위해 종단간 음성인식기의 각 부분별 동작을 시각적으로 이해하려고 하는 연구나, 기존의 음성인식 시스템처럼 별도로 훈련한 언어모델을 사용하여 성능을 개선하는 방법이 연구되고 있다(Gulcehre et al., 2015). 훈련 데이터의 양과 관련한 성능 연구로는 비교적 적은 양의 음성-문자열 짝 데이터를 이용하여 종단간 음성인식의 성능을 높이거나 음성이나 문자열만 존재하는 단일 데이터 또한 사용하여 음성인식기의 성능을 향상시키려는 연구가 진행되고 있다(Karita et al., 2018; Tjandra et al., 2017; Veselý et al., 2013). 본 연구에서는 음성 또는 문자열 단일 데이터를 함께 이용하여 종단간 음성인식기가 다른 도메인(domain)의 말뭉치에서도 좋은 성능을 낼 수 있도록 하는 준교사 학습 방식을 제안했으며, 각기 성격이 다른 도메인에 적용하여 제안된 방식이 어떤 경우 효과적인지 분석하였다.

1.2. 연구 목적

본 논문에서는 종단간 음성인식기를 위한 준교사 학습 방식의 도메인 적응 기법을 제안하며, 제안된 방식이 어떠한 경우 효과적으로 작동하는지 분석하였다. 앞서 기술한 바와 같이 종단간 음성인식기 학습에는 많은 양의 음성-문자열 짝 데이터가 필요하다. 하지만 이를 구하는 것은 사람의 노동력과 시간이 많이 필요하고, 좋은 성능의 종단간 음성인식기를 구축하는데 있어서 장벽이 되고 있다. 비교적 적은 양의 음성-문자열 짝 데이터와 음성 단일 데이터 또는 문자열 단일 데이터를 이용하여 학습하는 준교사 학습(semi-supervised training)으로 종단간 음성인식기의 성능을 개선하는 선행연구들이 있으나, 음성 단일 데이터나 문자열 단일 데이터 한쪽만을 사용하여 진행된 연구가 대부분이다. 본 연구에서는 기존 소스 데이터(source data)로 학습된 종단간 음성인식기를 이용하여 새로운 타깃 데이터(target data)에서도 좋은 성능을 낼 수 있도록 하는 준교사 학습 방식 도메인 적응 기법을 제안한다. 이 방식은 이미 학습된 종단간 음성인식기를 활용할 뿐만 아니라 음성 단일 데이터나 문자열 단일 데이터 양쪽을 모두 사용하는 방식이기 때문에 얻기 쉬운 다량의 데이터로 좋은 성능을 낼 수 있도록 함에 의의가 있다.

제안된 방식이 효과적인지 검증하기 위해 하나의 소스 도메인과 두 가지 다른 성격의 타깃 도메인을 활용하며, 타깃 도메인의 성격에 따라 기존의 다른 도메인 적응 방법과의 성능 차이를 비교한다. 또한 타깃 도메인의 적은 양의 음성-문자열 짝 데이터를 소스 도메인의 음성-문자열 짝 데이터와 병합하여 학습시킨 종단간 음성인식기와의 성능 비교를 통해 제안된 방식이 어떠한 점에서 더 유리한 방법인지 확인한다. 더불어 음성이나 문자열 단일 데이터의 양을 조절하며 적은 양의 단일 데이터만 있는 상황에서도 효과적으로 동작할 수 있는지 검증하고, 단일 데이터의 양을 충분히 가져갈 수 있으면 타깃 도메인에서의 음성-문자열 짝 데이터를 많이 확보하지 못해도 성능 향상을 기대할 수 있는지 확인하였다.

2. 준교사 학습 방식 도메인 적응

2.1. 사용 데이터

본 연구에서는 한 가지의 소스 데이터와 두 가지의 서로 다른 특징의 타깃 데이터를 이용하여 실험을 진행하였으며, 타깃 데이터의 특징에 따라 제안된 종단간 음성인식기를 위한 준교사 방식의 도메인 적응 방법이 다른 방법에 비해 효과적인지 확인할 수 있도록 하였다.

먼저 소스 데이터로는 LibriSpeech라는 레이블이 있는 음성 데이터를 사용하였다. 이 음성 데이터는 존스 홉킨스 대학에서 구성하여 무료로 배포하고 있으며, 무료 오디오북 프로젝트인 LibriVox의 영어 오디오북 데이터를 수집하여 데이터화 한 것이다. LibriSpeech에서 제공하는 데이터는 총 1,000시간 정도의 음성 데이터와 그에 상응하는 정답 레이블 문자 데이터를 제공한다. 본 실험에서는 종단간 음성인식기 역할을 하는 부분인 음성 인코더와 공유 디코더를 선행학습 시키기 위해 100시간 분량의 잡음이 없는 레이블이 있는 음성 데이터를 활용하였다. LibriSpeech 데이터의 녹음은 자연스러운 낭독체로 발화하여 녹음되었다.

첫 번째 타깃 데이터로는 Tedlium 2라는 데이터를 사용하였다. 이 음성 데이터는 Le Mans Universite의 LIUM에서 수집한 TED 연사 강연 녹음을 기반으로 만들어진 음성 데이터이며, 음성에 상응하는 정답 문자 레이블을 제공한다. 이 중에서 실험에 사용한 데이터로는 레이블이 있는 음성 데이터 40시간과 그에 상응하는 정답 레이블 문자열을 준교사 학습 도메인 적응에서 레이블이 있는 데이터로써 사용하였고, 음성 데이터 200시간과 그에 상응하는 정답 레이블 문자열은 레이블이 없는 데이터로써 활용하기 위해 순서를 섞어 각 음성과 문자열이 대응하지 못하도록 만들어 레이블이 없는 데이터로써 준교사 학습 시에 사용하였다. Tedlium 2 데이터에는 기본적으로 약간의 반향 잡음이 포함되어 있으며, 연사 강연의 녹음본인만큼 자유발화로 녹음되었다.

두 번째 타깃 데이터로는 WSJ 데이터를 사용하였다. 이 음성 데이터는 월스트리트 저널의 기사를 한 문장씩 전문 성우가 발화한 데이터로 이루어져 있으며 음성에 대응하는 정답 레이블 문자열을 제공한다. 이 중에서 실험에 사용한 데이터로는 레이블이 있는 음성 데이터 15시간과 그에 상응하는 정답 레이블 문자열을 준교사 학습 도메인 적응에서 레이블이 있는 데이터로써 사용하였고, 음성 데이터 80시간과 그에 대응하는 정답 레이블 문자열은 레이블이 없는 데이터로써 활용하기 위해 순서를 섞어 각 음성과 문자열이 대응하지 못하도록 만들어 레이블이 없는 데이터로써 준교사 학습 시에 사용하였다. WSJ 데이터의

녹음은 잡음이 없는 환경에서 진행되었으며, 뉴스 기사를 딱딱한 낭독체로 성우가 발화하여 녹음되었다. 특기할 사항으로는 이 데이터의 경우 기사에 등장하는 모든 문장 부호 또한 발화함으로써 실제 상황에서 발화되는 일상 언어와는 괴리가 있는 단어열로 발화되었다는 것이 특징이다.

2.2. 사용 툴킷

본 연구에서 준교사 학습 방식 도메인 적응을 위한 종단간 음성인식기를 구성하기 위해 사용한 툴킷은 ESPnet이다(Watanabe, 2018). ESPnet은 리눅스 기반 운영체제에서 작동하며 다양한 데이터에서 종단간 음성인식기와 음성합성기 모델을 학습할 수 있는 툴킷으로써 툴킷 내부에서 각 데이터별 최고 성능을 갖는 모델 또한 제공하고 있다. 이 툴킷은 종단간 음성 처리를 위한 Python 기반 오픈 소스 플랫폼으로 2018년 공개되었다. 하이브리드 CTC/어텐션 기반 손실 함수를 통한 음성인식과 Tacotron2 기반의 종단간 음성 합성, RNN-Transducer 기반의 종단간 음성인식 등 대다수의 종단간 모델을 지원하며 신경망 모델 구성에 주로 사용되는 Pytorch나 Chainer를 지원하여 연구자들이 새로운 모델 설계나 기존 모델의 확장을 쉽게 할 수 있도록 구성되었다.

2.3. 도메인 적응 모델의 구조

본 논문에서 제안하는 모델의 구조는 그림 1과 같다. 제안된 모델은 크게 세 부분으로 나누어져 있다. 음성 인코더(speech encoder), 문자 인코더(text encoder), 그리고 공유 디코더(shared decoder)로 구성되어 있다. 먼저 음성 인코더와 공유 디코더 두 부분은 함께 동작하여 어텐션 기반의 인코더-디코더 종단간 음성인식기와 같은 동작을 한다. 음성 인코더의 입력으로 레이블이 있는 음성 신호가 들어가며 인코더를 통해 음성 은닉 벡터로 변환된다. 음성 은닉 벡터는 다시 공유 디코더의 입력으로 들어가며 어텐션과 함께 동작하여 음성 인코더의 입력 음성 신호의 레이블에 대응하는 정답 문자열을 출력하게 된다. 다음으로는 문자 인코더와 공유 디코더가 같이 동작하여 텍스트 오토인코더(text auto-encoder)처럼 동작한다. 이는 문자열만을 입력으로 받으며 문자 인코더에서 문자 은닉 벡터로 변환된 후 공유 디코더에 들어가서 어텐션과 함께 작동해 최종 출력으로는 입력되었던 문자와 동일한 문자열을 출력하게 된다. 즉 제안된 모델은 음성인식기와 텍스트 오토인코더를 결합해 놓은 모델로 생각할 수 있다. 이 모델이 학습될 때 레이블이 없는 문자열 단일 데이터의 정보가 이용되고, 이는 공유 디코더와 어텐션 부분이 학습될 때 영향을 주게 된다.

음성 인코더와 공유 디코더가 동작하여 음성 입력을 받아 정답 문자 출력을 하는 부분은 기존 종단간 음성인식기와 같은 손실 함수를 사용하며(L_{asr}) 문자열 입력을 받아 그와 똑같은 문자열을 출력하도록 학습되는 문자 인코더와 공유 디코더 부분의 학습에는 텍스트 오토인코더 손실 함수가 사용된다(L_{tae}). 또한 음성 인코더가 출력하는 음성 은닉 벡터와 문자 인코더가 출력하는 문자 은닉 벡터가 유사한 분포 공간상에 있도록 유도하는 손실 함수를 두어 어텐션과 공유 디코더 부분이 학습 과정에서 음성과 문자열 정보의 차이를 알 수 어렵게 하여 공통된 정보를 더 잘 학습할 수 있도록 유도하였다. 이 손실 함수는 인터모달리티 손실(intermodality loss, L_{mod}) 로 정의하였다.

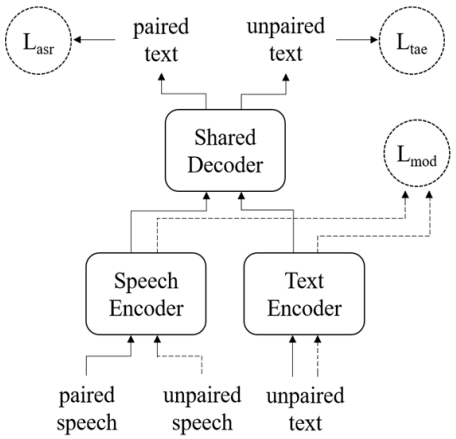

그림 1. 준교사 학습 방식의 도메인 적응 종단간 음성인식 모델
Figure 1. Semi-supervised domain adaptation end-to-end automatic speech recognition model

x와 y는 각각 레이블이 있는 음성과 문자 벡터를 나타내며, x'와 y'는 레이블이 없는 음성과 문자 벡터를 나타낸다. 공유 디코더와 인터모달리티 손실은 제안된 모델에서 가장 중요한 역할을 하는 부분이다. 각 인코더에 입력으로 주어지는 음성 신호와 문자열은 각각 상이한 길이와 값들을 갖고 있다. 따라서 음성 인코더와 문자 인코더가 출력하는 은닉 벡터는 모종의 조치 없이는 서로 다른 분포를 갖게 된다. 그렇기에 음성 인코더는 길이 보상을 통해 문자 인코더의 출력과 비슷한 길이로 은닉 벡터를 출력할 수 있도록 조절하며, 문자 인코더는 CNN 기반 레이어와 LSTM을 통해 이산적인 신호인 문자열 벡터를 실제 음성 신호와 비슷한 연속적인 벡터로 변환하게 된다. 인터모달리티 손실은 음성 인코더와 문자 인코더가 출력하는 은닉 벡터가 비슷한 분포를 따르도록 학습될 수 있도록 도와준다. 이 손실 함수는 Kullback-Leibler 발산으로 계산된다. 인터모달리티 손실 함수를 적용함으로써 음성 또는 문자열 단일 데이터를 다량 활용하여 음성과 문자열 사이의 상호 관계를 모델 학습 과정에서 배울 수 있게 되며, 이는 기존 준교사 방식들이 가졌던 한계인 음성과 문자열 단일 데이터를 동시에 사용할 수 없다는 문제점을 해결한다.

제안된 모델의 각 부분의 수치는 다음과 같다. 먼저 음성 인코더의 경우 VGG-CNN 구조의 임베딩 레이어(embedding layer)와 5층의 BLSTM으로 이루어져 있으며, 각 레이어는 300차의 투영 벡터를 출력으로 한다. 문자 인코더의 경우도 음성 인코더와 비슷한 구조를 가지며 5×5 크기의 필터를 갖는 CNN 레이어 1층과 더불어 2층의 BLSTM 레이어로 이루어져 있으며 BLSTM은 300차의 투영 벡터를 출력으로 하여 음성 인코더에서 출력

되는 은닉 벡터와 같은 길이의 출력을 낸다. 어텐션 부분은 위치 기반의 어텐션을 사용하며 300차의 투영 벡터를 갖는다. 마지막으로 디코더는 1층의 LSTM으로 구성되어 있으며 300차의 투영 벡터를 출력으로 갖는다. 더불어 인코더의 입력 log-Mel filterbank 벡터는 80차원의 벡터를 음성 신호로부터 추출하여 사용했으며, 디코더의 출력은 30개의 문자 토큰(a-z, ‘, <blank>, <unk>, <eos>) 을 인식하도록 설정되었다.

종단간 모델에서 훈련 및 인식에 사용하는 문자 토큰은 CTC 기반의 문자 변형과 같은 원리로 변형된다. 문장의 처음에 “_” 토큰이 삽입되며, 문장의 끝에는 “<eos>” 토큰이 삽입된다. 또한 각 알파벳은 그에 대응하는 알파벳 토큰으로, 공백은 “_” 토큰으로 대체한다. 또한 아포스트로피를 제외한 모든 기호들은 제거하게 된다. 이렇게 생성된 토큰 문자열을 정답으로 하여 모델의 훈련에 사용하고, 인식 결과로 생성된 토큰 형태의 문장을 규칙에 따라 역변환 하여 정답과 비교하게 된다.

2.4. 도메인 적응 방법

제안된 모델의 학습은 크게 선행학습과 도메인 적응 두 단계로 이루어진다. 먼저 선행학습은 소스 도메인의 레이블이 있는 데이터로부터 음성 인코더와 공유 디코더를 학습시켜 기본적인 종단간 음성인식기처럼 동작하도록 한다. 이렇게 생성된 초깃값을 이용하여 도메인 적응을 진행하게 된다. 도메인 적응 단계에서는 앞서 생성된 음성 인코더와 공유 디코더의 초깃값을 이용해 타깃 도메인의 데이터로 재학습을 진행하게 된다. 타깃 도메인의 데이터는 비교적 적은 양의 레이블이 있는 데이터와 비교적 많은 양의 레이블이 없는 데이터로 이루어져 있다. 이 단계에서는 총 세 가지의 손실 함수가 사용되며, 이 손실 함수들이 구해지는 방식은 다음과 같다.

$$L_{asr} = -\log \Pr(y|dec(enc_{speech}(x))) \quad (1)$$

$$L_{tae} = -\log \Pr(y'|dec(enc_{text}(y'))) \quad (2)$$

$$L_{mod} = KL(enc_{speech}(x')\|enc_{text}(y')) \quad (3)$$

첫 번째로 L_{asr}가 구해진다. 이는 타깃 데이터의 레이블이 있는 데이터로만 구해지는 손실이며, 레이블이 있는 음성 입력이 음성 인코더와 공유 디코더를 거쳐 출력되는 문자열과 그 음성 입력의 정답 레이블을 비교하여 구해지는 손실이다. 이 손실만 사용할 경우 보통의 종단간 음성인식기의 재학습을 데이터 도메인만 바꾸어 진행한 파인 튜닝(fine tuning)으로 간주될 수 있다. 두 번째 손실로는 L_{tae}가 구해진다. 이는 타깃 데이터의 문자 단일 데이터만을 가지고 학습되며, 문자 입력이 문자 인코더와 공유 디코더를 거쳐 출력되는 문자열을 입력 문자열과 비교하여 구해지는 손실이다. 이 손실은 비교적 많은 양의 문자 단일 데이터의 정보를 디코더와 어텐션이 학습할 수 있도록 도와주는 역할을 하게 된다. 마지막으로 L_{mod}가 구해진다. 이는 타깃 데이터의 레이블이 없는 음성 단일 데이터와 문자 단일 데이터로 구해지는 손실이다. 이 손실은 미니배치(mini batch) 단위에서의 분포를 통해 계산되는 값이며, 공유 디코더가 음성 은닉 벡터와 문자 은닉 벡터 사이의 차이점을 알 수 없게 하여 종단간 음성인식기가 텍스트 오토인코더와의 병렬 학습을 더 효과적으로 이용할 수 있게 해 주는 손실이다. 이 손실을 구할 때 음성 단일 데이터와 문자 단일 데이터는 서로 의미가 다른 집합들로 이루어져 있으나, 다량의 데이터를 사용함으로써 음성과 문자가 각각 인코딩된 벡터가 같은 공간상에 분포하도록 학습할 수 있다.

위와 같이 구해진 세 손실은 각 미니배치마다 계산되어 다음과 같은 트레이닝 로스가 계산된다.

$$L = (1-\alpha)L_{asr} + \alpha[(1-\beta)L_{tae} + \beta L_{mod}] \quad (4)$$

매개변수는 각각 0.1–0.9 사이에서 조정하며 최적의 값을 구하였으며, 이를 구하는 데에는 검증 데이터(validation data)를 이용되었다.

2.5. 타깃 도메인별 도메인 적응

먼저, 소스 데이터로 LibriSpeech를 사용하고 타깃 데이터로는 Tedlium 2와 WSJ 데이터를 각각 사용하여 LibriSpeech에서 Tedlium 2로의 도메인 적응과 LibriSpeech에서 WSJ로의 도메인 적응 성능을 비교하였다. 제안된 모델과 비교할 모델을 두 가지 구성하였다. 소스 데이터와 타깃 데이터의 레이블이 있는 데이터를 모두 활용해 학습시킨 종단간 음성인식 모델과 소스 데이터로 선행학습된 종단간 음성인식기를 타깃 데이터로 재학습시킨 모델을 제안한 모델과 비교하였다. 또한 소스 도메인에서 가장 좋은 성능을 낼 소스 데이터의 레이블이 있는 데이터로만 학습된 종단간 음성인식기와 타깃 도메인에서 가장 좋은 성능을 낼 타깃 데이터의 레이블이 있는 데이터로만 학습된 종단간 음성인식기도 각각 훈련하여 성능을 비교하였다. 성능 비교는 모두 문자 오인식율(character error rate, CER)로 비교하였다. 또한 적응 지표라는 수치를 제안하여 사용하였다. 적응 지표의 계산은 다음과 같다.

$$\text{적응 지표, \%} := \text{타깃 향상율} - \text{소스 열화율} \quad (5)$$

타깃 향상율은 타깃 도메인에서 모델의 성능이 얼마나 최대치에 근접하는지를 나타내며, 타깃 도메인에서의 CER 수치들로 다음과 같이 구해진다. 먼저 적응 지표를 계산할 모델의 타깃 도메인 CER을 M_T, 타깃 도메인의 데이터로만 학습시킨 모델의 타깃 도메인 CER을 T_T, 소스 도메인의 데이터로만 학습시킨 모델의 타깃 도메인 CER을 S_T라고 하면,

$$\text{타깃 향상율, \%} := (S_T - M_T) \div (S_T - T_T) \quad (6)$$

소스 열화율은 소스 도메인에서 모델의 성능이 얼마나 최저치에 근접하는지를 나타내며, 소스 도메인에서 CER 수치들로 다음과 같이 구해진다. 타깃 향상율을 구하는 경우와 비슷하게

적응 지표를 계산할 모델의 소스 도메인 CER을 M_S, 타깃 도메인의 데이터로만 학습시킨 모델의 소스 도메인 CER을 T_S, 소스 도메인의 데이터로만 학습시킨 모델의 소스 도메인 CER을 S_S 라고 하면,

$$\text{소스 열화율, \%} := (M_S - S_S) \div (T_S - S_S) \quad (7)$$

로 계산된다.

즉, 적응 지표의 수치가 클수록 타깃 도메인에서의 성능 향상은 크고 소스 도메인에서의 성능 열화는 적은 모델이 된다.

2.6. 단일 데이터의 양에 따른 도메인 적응

다음은 타깃 도메인에서 사용할 수 있는 레이블이 있는 데이터는 한정되어 있고, 타깃 도메인의 레이블이 없는 데이터의 양을 바꾸어 가며 실험을 진행하였다. 소스 도메인의 데이터로는 LibriSpeech 데이터를 활용하였고, 타깃 도메인의 데이터로는 Tedlium 2의 데이터를 사용하였다. 모든 모델은 공통적으로 타깃 도메인의 레이블이 있는 데이터는 40시간가량의 데이터를 사용했으며, 타깃 도메인의 레이블이 없는 데이터의 양을 200시간 가량 확보하고, 그 양을 25%–100%까지 조절해 가며 실험을 진행하였다.

3. 실험 결과

다음 표 1과 표 2는 각 타깃 도메인 데이터로의 도메인 적응 성능을 모델별로 비교한 것이다.

표 1. LibriSpeech 소스 도메인에서 Tedlium 2 타깃 도메인으로의 적응 % 성능 비교

Table 1. The measure of domain adaptation model CER% and adaptation indicator % from LibriSpeech source domain to Tedlium 2 target domain

모델	소스 CER	타깃 CER	적응 지표
소스로만 학습	6.8	21.5	-
모든 데이터로 학습	9.8	18.5	–4.1
제안된 모델	13.9	12.2	+10.6
재학습	14.9	13.9	–6.4
타깃으로만 학습	16.3	10.6	-

CER, character error rate.

표 2. LibriSpeech 소스 도메인에서 WSJ 타깃 도메인으로의 적응 % 성능 비교

Table 2. The measure of domain adaptation model CER% and adaptation indicator % from LibriSpeech source domain to WSJ target domain

모델	소스 CER	타깃 CER	적응 지표
소스로만 학습	6.8	12.9	-
모든 데이터로 학습	7.5	12.2	+6.2
제안된 모델	13.0	9.4	+13.3
재학습	19.2	8.1	–7.7
타깃으로만 학습	21.8	6.5	-

CER, character error rate.

먼저 표 1에서 볼 수 있듯이 타깃 도메인의 성능에서 가장 좋은 성능을 보이는 것은 타깃 도메인의 데이터로만 학습된 모델이다. 이에 가장 근접한 성능을 보이는 모델은 제안된 모델이고, 준교사 방식 도메인 적응 기법이 잘 동작하는 것을 볼 수 있다. 제안된 모델과 단순 재학습한 모델의 차이점은 제안된 모델이 타깃 데이터의 레이블이 없는 데이터를 학습에 사용했다는 점이다. 이로 인해 레이블이 없는 타깃 도메인의 데이터로부터 음성인식에 필요한 모종의 정보를 더 학습 가능했으며 결과적으로 타깃 도메인에서의 성능 향상에 기여했다고 볼 수 있다. 또한 적응 지표에서 나타나듯이 제안된 모델이 다른 두 모델과 달리 소스 도메인에서의 성능 열화가 적고 타깃 도메인에서의 성능 향상이 큰 것으로 나타났다. 이는 인터모달리티 손실 함수를 통해 타깃 도메인의 음성과 문자 사이의 분포를 학습하여 공유 디코더가 올바른 방향으로 학습된다는 것을 보여준다. 반면 재학습된 모델의 경우 타깃 도메인에서의 성능 향상은 적지 않으나 소스 도메인에서의 성능 열화가 비교적 더 크게 나타났고, 이는 소스 도메인과 타깃 도메인의 차이점에 의해 나타나는 현상으로 분석된다. 또한 모든 도메인의 레이블이 있는 데이터를 함께 사용하여 학습시킨 모델의 경우 재학습된 모델보다는 적응 지표가 높게 나타났으나, 타깃 도메인에서 사용할 수 있는 레이블이 있는 데이터의 양이 제한되어 있기 때문에 소스 도메인의 데이터로만 학습된 모델에서 크게 변화하지 못하는 성능을 보였고, 이는 실제 상황의 적용에도 타깃 도메인의 레이블이 있는 데이터를 많이 구하지 못하는 상황과 비슷한 상황이 된다. 타깃 도메인의 레이블이 없는 데이터를 구하는 것은 레이블이 있는 데이터를 구하는 것보다 손쉬우며, 제안된 모델은 레이블이 없는 타깃 도메인의 데이터를 함께 활용함으로써 타깃 도메인으로의 적응을 손쉽게 가능하도록 했다는 점에서 효과가 크다고 분석된다.

표 2에서의 성능 수치는 표 1과는 조금 다른 양상을 보인다. 마찬가지로 제안된 모델 또한 타깃 도메인에서의 성능 향상이 크게 나타난다. 하지만 타깃 도메인에서의 성능 향상이 가장 큰 모델은 타깃 도메인의 데이터로 재학습된 모델인 것으로 나타난다. 반면 적응 지표를 보면 제안된 모델의 경우 소스 데이터에서의 열화가 크지 않고 타깃 도메인에서 성능 향상이 많이 되었지만 타깃 도메인의 데이터로 재학습된 모델의 경우 적응 지표에서 볼 수 있듯이 소스 데이터에서의 열화가 매우 큰 것으로 나타난다. 이는 소스 도메인인 LibriSpeech와 타깃 도메인인 WSJ의 성격이 매우 다름에서 기인한 것이다. WSJ 데이터는 LibriSpeech와 유사하게 낭독체로 발화된 음성이나, WSJ 데이터의 발화는 문장기호를 모두 발화하기에 현실적인 문장 구조에서의 단어 나열과는 큰 차이점을 보인다. 종단간 음성인식기가 훈련되는 과정에서 어텐션 기반의 인코더-디코더 구조를 통해 음성과 문자의 정렬을 자체적으로 학습하며, 이는 학습 데이터의 문장 구조에도 큰 영향을 받게 된다. 따라서 단순 재학습된 모델의 경우 소스 도메인의 문장 구조를 벗어나 타깃 도메인의 문장 구조만을 학습하는 방향으로 모델이 결정되며, 이는 타깃 도메인에서의 성능은 높일지라도 소스 도메인에서의 성능

을 크게 열화시킬 수밖에 없는 것으로 분석된다. 이와 다르게 제안된 모델은 인터모달리티 손실 함수를 통해 타깃 도메인에서의 단어의 배열과 음성 프레임의 관계를 학습함으로써 소스 도메인에서 나타났던 문장 구조를 크게 손실시키지 않고 보다 올바른 방향의 음성인식기로써 학습이 가능하게 된다. 표 1과 비슷하게 표 2에서도 모든 데이터로 학습한 모델의 성능은 소스 도메인의 데이터로만 학습한 모델의 성능과 크게 다르지 않음이 나타나며, WSJ의 데이터의 양이 표 1에서 타깃 데이터로 사용한 Tedlium 2의 데이터의 양보다 더 적기 때문에 이와 같은 현상이 더 크게 나타나는 것으로 분석된다.

이 실험에서 알 수 있듯이 제안된 준교사 방식 도메인 적응 기법은 소스 도메인으로 학습된 기존 모델을 타깃 도메인의 적은 양의 레이블 있는 데이터와 비교적 구하기 쉬운 타깃 도메인의 많은 양의 레이블이 없는 데이터를 함께 이용하여 타깃 도메인에서의 성능을 향상시킬 수 있다는 장점이 있다. 또한 단순 재학습시 발생하는 소스 도메인에서의 음성인식기 성능 열화를 억제하고 타깃 도메인으로 과적합될 수 있는 문제를 해결한 점에서 의의가 있다.

다음 표 3은 단일 데이터의 양에 따른 도메인 적응 성능을 나타낸 것이다.

표 3. 단일 데이터의 양에 따른 적응 % 성능 비교
Table 3. The measure of domain adaptation model CER% and adaptation indicator % based on the amount of unlabeled data

모델	소스 CER	타깃 CER	적응 지표
소스로만 학습	6.8	21.5	-
모든 데이터로 학습	9.8	18.5	–4.1
25%	8.5	18.8	+6.9
50%	10.5	16.4	+7.8
75%	12.1	14.5	+8.4
100%	13.9	12.2	+10.6
재학습	19.2	8.1	–7.7

표 3에 나타난 수치에서 볼 수 있듯이 단일 데이터의 양을 많이 활용할수록 타깃 도메인에서의 성능이 향상되는 것을 볼 수 있었다. 이는 준교사 학습 방식이 종단간 음성인식기의 도메인 적응 과정에서 효과적이라는 것을 나타낸다. 적은 양의 단일 데이터만을 활용한 경우에도 위에서 비교한 소스 도메인과 타깃 도메인의 모든 데이터로 학습한 모델과 비교하여 타깃 도메인에서의 성능은 비슷한 수치를 보이지만 소스 도메인에서의 성능 하락은 더 적은 것으로 나타났다. 이는 제안된 방법이 적은 양의 단일 데이터를 사용하더라도 전혀 사용하지 않고 학습한 경우보다 소스 도메인에서의 성능 하락을 줄일 수 있는 방법임을 의미한다. 또한 주목할 만한 변화로는 단일 데이터의 양을 늘려감에 따른 적응 지표의 증가가 있다. 실험에서 사용된 단일 데이터의 양을 늘려감에 따라 적응 지표 또한 증가하는 것을 볼 수 있다. 적응 지표의 향상은 단일 데이터의 양이 레이블이 있는 데이터의 양보다 과도하게 많지 않을 때까지는 향상할 것으로 보이며, 한 수치로 수렴할 것으로 사려된다. 이는 제안된 모델에서 인터모달리티 손실 함수의 도입이 의도한 결과이며 레이블이 없는 음성과 문자열일지라도 다수 확보될수록 종단간 음성인식기의 도메인 적응 성능 향상을 가능하게 한다는 점을 보여 준다. 결과적으로 종단간 음성인식기를 특정 도메인에서 성능을 향상시키고자 할 때 레이블이 있는 데이터만을 많이 확보하지 않아도 비교적 빠르고 얻기 쉬운 음성 데이터만을 다수 확보하고 타깃 도메인과 성격이 유사한 말뭉치의 문자 데이터를 이용하여 적은 시간과 비용으로 타깃 도메인에서 좋은 성능을 보이는 종단간 음성인식기를 구축할 수 있다. 또한 제안된 방식은 기존의 종단간 음성인식기 구조를 그대로 가져가면서 문자 인코더와 인터모달리티 손실 함수만 추가함으로써 학습 시간 또한 기존 종단간 음성인식기를 학습하는 것과 큰 차이 없이 수행할 수 있어 경제적이다.

4. 결론

본 연구에서는 종단간 음성인식기를 위한 준교사 학습 방식의 도메인 적응을 진행하였고, 이를 통해 레이블이 있는 타깃 데이터의 양이 불충분하더라도 비교적 얻기 쉬운 타깃 데이터의 음성 단일 데이터나 문자열 단일 데이터를 이용하여 효과적으로 타깃 도메인으로 적응할 수 있는 기법을 제안하였다. 현재 종단간 음성인식기가 만족스러운 성능을 내기 위해서는 기존의 고전적인 음성인식기의 훈련에 사용되는 음성-문자열 짝 데이터의 양보다 많은 양의 데이터를 확보해야 하는 것이 종단간 음성인식기를 구축함에 있어 큰 진입장벽이 되고 있었다. 이러한 문제를 해결하기 위해 시간과 비용이 덜 필요한 레이블이 없는 단일 데이터를 활용함으로써 종단간 음성인식기가 다른 도메인에서도 효과적으로 동작할 수 있는 방법을 제안하였다.

본 연구에서는 어텐션 기반 인코더-디코더 구조를 갖는 종단간 음성인식기에 문자 인코더를 추가하고 각 인코더가 출력하는 은닉 음성 벡터와 은닉 문자 벡터가 유사한 공간에 분포하도록 하는 인터모달리티 손실 함수를 도입한 도메인 적응 모델을 제안하고 훈련하였다. 다른 성격을 갖는 타깃 도메인에서 유효한 성능을 발휘하는지 확인하기 위해 LibriSpeech 데이터를 소스 도메인으로 이용하고 타깃 도메인으로는 Tedlium 2와 WSJ 데이터를 사용하여 각각 도메인 적응 실험을 진행하였다. 두 도메인으로의 적응 실험 모두 기존의 재학습 기법보다 좋은 성능 또는 비슷한 성능을 냄으로써 제안된 방식이 종단간 음성인식기의 도메인 적응에 사용될 수 있음을 확인하였다. 또한 제안된 방식이 타깃 도메인에서의 성능 향상과 더불어 소스 도메인에서의 성능 열화를 적게 한다는 장점 또한 적응 지표를 통해 확인할 수 있었다. 추가적으로 단일 데이터의 양을 달리해 실험을 진행하여 본 연구에서 제안된 모델이 비교적 적은 양의 단일 데이터의 상황에서도 도메인 적응 상황에 유효한지 확인하였으며, 다량의 단일 데이터가 확보되었을 경우 좋은 성능의 도메인 적응 기법으로써 작용할 수 있다는 것을 확인하였다.

References

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). Shanghai, China.

Graves, A., Fernández, 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). Pittsburgh, PA.

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). Vancouver, Canada.

Gulcehre, C., Firat, O., Xu, K., Cho, K., Barrault, L., Lin, H., Bougares, F., … Bengio, Y. (2015, June). On using monolingual corpora in neural machine translation [Computing research repository]. Retrieved from https://arxiv.org/pdf/1503.03535.pdf

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., Senior, A., … Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, *29*(6), 82-97.

Karita, S., Watanabe, S., Iwata, T., Ogawa, A., & Delcroix, M. (2018, September). Semi-supervised end-to-end speech recognition. *Proceedings of the International Conference on Spoken Language Processing (INTERSPEECH)* (pp. 2-6). Taipei, Taiwan.

Miao, Y., Gowayyed, M., & Metez, 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). Scottsdale, AZ.

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

Tjandra, A., Sakti, S., & Nakamura, S. (2017, December). Listening while speaking: Speech chain by deep learning. *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (pp. 301-308). Okinawa, Japan.

Veselý, K., Hannemann, M., & Burget, L. (2013, December). Semi-supervised training of deep neural networks. *IEEE Workshop on Automatic Speech Recognition and Understanding* (pp. 267-272). Olomouc, Czech.

Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Soplin, N. E. Y., … Ochiai, T. (2018). ESPnet: End-to-end speech processing toolkit [Computing research repository]. Retrieved from http://arxiv.org/abs/1804.00015

- **정현재 (Hyeonjae Jeong)**
  한국과학기술원 전기및전자공학부 석사
  대전광역시 유성구 대학로 291
  Tel: 042-350-7617
  Email: kkkwjd12@kaist.ac.kr
  관심 분야: 음성인식

- **구자현 (Jahyun Goo)**
  한국과학기술원 전기및전자공학부 박사과정
  대전광역시 유성구 대학로 291
  Tel: 042-350-7617
  Email: jahyun.goo@kaist.ac.kr
  관심 분야: 음성인식

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

라벨이 없는 데이터를 사용한 종단간 음성인식기의 준교사 방식 도메인 적응

정 현 재 · 구 자 현 · 김 회 린

한국과학기술원 전기및전자공학부

국문초록

최근 신경망 기반 심층학습 알고리즘의 적용으로 고전적인 Gaussian mixture model based hidden Markov model (GMM-HMM) 음성인식기에 비해 성능이 비약적으로 향상되었다. 또한 심층학습 기법의 장점을 더욱 잘 활용하는 방법으로 언어모델링 및 디코딩 과정을 통합처리 하는 종단간 음성인식 시스템에 대한 연구가 매우 활발히 진행되고 있다. 일반적으로 종단간 음성인식 시스템은 어텐션을 사용한 여러 층의 인코더-디코더 구조로 이루어져 있다. 때문에 종단간 음성인식 시스템이 충분히 좋은 성능을 내기 위해서는 많은 양의 음성과 문자열이 함께 있는 데이터가 필요하다. 음성-문자열 짝 데이터를 구하기 위해서는 사람의 노동력과 시간이 많이 필요하여 종단간 음성인식기를 구축하는 데 있어서 높은 장벽이 되고 있다. 그렇기에 비교적 적은 양의 음성-문자열 짝 데이터를 이용하여 종단간 음성인식기의 성능을 향상하는 선행연구들이 있으나, 음성 단일 데이터나 문자열 단일 데이터 한쪽만을 활용하여 진행된 연구가 대부분이다. 본 연구에서는 음성 또는 문자열 단일 데이터를 함께 이용하여 종단간 음성인식기가 다른 도메인의 말뭉치에서도 좋은 성능을 낼 수 있도록 하는 준교사 학습 방식을 제안했으며, 성격이 다른 도메인에 적응하여 제안된 방식이 효과적으로 동작하는지 확인하였다. 그 결과로 제안된 방식이 타깃 도메인에서 좋은 성능을 보임과 동시에 소스 도메인에서도 크게 열화되지 않는 성능을 보임을 알 수 있었다.

핵심어: 음성인식, 종단간, 준교사 학습, 도메인 적응