pISSN 2005-8063
eISSN 2586-5854
2020. 12. 31.
Vol.12 No.4
pp. 63-71

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2020.12.4.063

End-to-end speech recognition models using limited training data*

June-Woo Kim · Ho-Young Jung**

Department of Artificial Intelligence, Kyungpook National University, Daegu, Korea

Abstract

Speech recognition is one of the areas actively commercialized using deep learning and machine learning techniques. However, the majority of speech recognition systems on the market are developed on data with limited diversity of speakers and tend to perform well on typical adult speakers only. This is because most of the speech recognition models are generally learned using a speech database obtained from adult males and females. This tends to cause problems in recognizing the speech of the elderly, children and people with dialects well. To solve these problems, it may be necessary to retain big database or to collect a data for applying a speaker adaptation. However, this paper proposes that a new end-to-end speech recognition method consists of an acoustic augmented recurrent encoder and a transformer decoder with linguistic prediction. The proposed method can bring about the reliable performance of acoustic and language models in limited data conditions. The proposed method was evaluated to recognize Korean elderly and children speech with limited amount of training data and showed the better performance compared of a conventional method.

Keywords: speech recognition, end-to-end model, small-data speech recognition

1. 서론

음성 인식은 인공지능 기술을 이용하는 서비스 가운데 활발히 상용화되고 있는 분야 중 하나이다. 음성 인식의 목적은 입력된 음성 신호로부터 텍스트를 얻는 것으로 음성 인식 시스템은 화자 종속과 화자 독립 유형으로 구분된다. 화자 종속은 한 사람 목소리의 고유한 개별 특성을 학습함으로써 작동한다. 사용자는 음성 인식 시스템을 활용하기 위해서 사전 학습을 할 필요가 있다. 새로운 사용자가 화자 종속 음성 인식 시스템을 사용하기 위해서는 미리 해당 사용자의 데이터를 학습해야 하며, 이것은 소규모 어휘를 인식하는 경우에 적합하다. 이에 비해 화자 독립 음성 인식 시스템은 불특정 다수 화자의 음성을 인식하도록 개발되는 것이다. 따라서, 다양한 음성 인식 서비스를 사용자에 무관하게 제공하기 위해서는 화자 독립 음성 인식 시스템을 구축하는 것이 요구된다.

현재 활발히 연구되고 있는 새로운 음성 인식 기술과 국내외 서비스 중인 음성 인식 시스템은 대부분 성인 남녀의 음성을 상대적으로 잘 인식하는 편이다. 이는 음성 인식 모델이 성인 남

* This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (2020-0-01808, Research on innovative prediction intelligence technology using multimodal information).

** hoyjung@knu.ac.kr, Corresponding author

Received 31 July 2020; Revised 16 November 2020; Accepted 16 November 2020

녀 기준의 음성 데이터베이스를 이용하여 학습이 이루어지고 있기 때문이다. 현재 음성 인식 기술 연구를 위해 많이 사용되고 있는 LibriSpeech 데이터베이스는 2,484명의 화자가 약 1,000시간의 양을 녹음한 음성 데이터이며, 대부분 20–50대의 화자들로 이루어져 있다.

국내에서 연구 목적으로 공개한 음성 데이터베이스도 대개 성인 남녀 기준으로 구성되어 있으며, 음성 인식 서비스를 제공하는 기업에서도 데이터 수집의 어려움과 비용 등의 문제로 성인 남녀의 음성을 중심으로 음성 인식 엔진을 개발하고 있는 실정이다. 따라서, 노인 및 어린이 음성 인식을 위한 시스템 개발에 많은 어려움이 존재하고 있다. 이들의 음성 특징은 성인 남녀와 다르며, 특히 자유 발화 노인 음성의 경우 성인 남녀 음성으로 학습된 음성 인식 엔진을 적용하면 성능이 급격히 떨어지는 문제가 있다. 노인과 어린이의 음성 인식 성능을 높이기 위해서는 빅데이터를 구축하는 방법과 성인 대상 음성 인식 엔진을 노인 및 어린이 데이터로 적응하는 방법이 있으나 추가적인 비용과 시간이 요구된다. 이 문제를 해결하기 위해서는 제한된 데이터로 음성 인식 성능을 개선하기 위한 새로운 모델 구조가 필요할 것이다.

본 논문에서는 위의 방법 가운데 제한된 데이터를 이용하여 강인한 음성 인식 학습이 가능한 end-to-end 모델을 제안하고, 노인 및 어린이 데이터를 이용하여 성능을 평가한다. 제안된 모델은 음향적 정보를 증강하는 재귀적 뉴럴 네트워크 기반의 인코더와 언어적 변이의 다양성을 학습하도록 개선된 Transformer (Vaswani et al., 2017) 디코더로 이루어진 end-to-end 방식이다. 인코더는 제한된 학습 데이터의 문제를 해결하여 다양한 음향적 특성을 반영할 수 있고, 디코더는 학습 데이터의 텍스트 정보로만 이루어지는 제한적인 언어모델의 문제를 해결할 수 있을 것으로 기대된다. 어린이 및 노인 음성 데이터베이스는 공개된 데이터를 활용하였다.

본 논문에서 사용한 공개 노인 음성 데이터베이스는 마인즈랩과 한국전자통신연구원이 협력하여 구축한 VOTE400(Voices Of The Elders 400 Hours) 데이터셋이었고, 어린이 데이터베이스는 한국전자통신연구원에서 구축하여 공개한 초등학생 음성 데이터셋 및 엘솔루에서 공개한 어린이 음성 데이터셋이었다. 노인 음성 데이터셋은 어린이 음성 데이터셋에 비해 상대적으로 용량이 크므로, 전체 학습 데이터에서 일부분을 분리하고 학습 데이터 증가 전후의 성능 평가도 같이 진행하여 제한된 데이터 조건에서 제안된 방법이 효과적으로 모델을 학습함을 보여준다.

본 논문은 2장에서 기존 end-to-end 음성 인식 엔진에 대해 살펴보고, 3장에서 음향적 정보를 증강하는 재귀적 뉴럴 네트워크 기반의 인코더와 언어적 다양성을 학습하는 개선된 Transformer 디코더로 이루어진 end-to-end 구조에 대해 설명한다. 4장에서는 음성 인식 실험 환경 및 평가 결과를 설명하며, 5장에서 결론을 맺는다.

2. 기존 end-to-end 음성 인식 모델

인코더-디코더 형태를 가지는 sequence-to-sequence(Sutskever et al., 2014) 기반 초기 end-to-end 음성 인식 모델인 LAS(Listen, Attend, and Spell)에 대해 소개하고, 이와 더불어 최근 널리 활용되고 있는 ESPnet(Watanabe et al., 2018) 음성 인식 모델을 다루는 것을 통해 제안된 모델과 비교한다.

2.1. LAS: Listen, attend and spell

LAS(Chan et al., 2016) 모델은 기존 DNN-HMM(deep neural network-hidden Markov model) (Hinton et al., 2012) 구조 또는 CNN(convolutional neural network)-HMM 구조와 다르게 발성된 음성 신호로부터 텍스트를 바로 출력하는 신경망이다. LAS 모델이 기존 방법과 다른 점은 HMM에서 정의하는 음소 심벌 및 음소별 상태 분포 등의 정보를 전혀 필요로 하지 않으면서 단지 음성-텍스트 쌍으로 end-to-end 학습할 수 있다는 것이다. LAS 모델이 제안될 당시 음성 인식 모델은 음향모델, 언어모델, 발음 사전, 텍스트 정규화 등 다양한 구성 요소로 이루어진 복잡한 시스템이었다. 언어모델과 HMM 구조는 문장 내 단어 사이의 연관 관계 및 단어-음소 기반 음성 데이터의 상태 모델과 관측 확률 분포 모델을 만든다. 이로 인해 DNN-HMM 시스템은 심층 신경망이 관측 확률 분포를 프레임 세그먼트에 기반하여 추정할 수 있도록 하지만, 음소 단위 상태 천이 모델과 N-gram 기반의 언어모델을 이용하여 시간 경과에 따른 예측 모델을 학습하는 별도 과정을 요구한다.

End-to-end 음성 인식 이전에는 위에서 보듯이 2개 이상의 구성 요소를 합쳐서 음성 인식 시스템을 만들었지만, LAS는 다른 구성 요소는 배제하고 음성을 바로 문자로 변환하는 모델의 구현을 시도하였다. 즉, 모델의 입력으로 음성 신호를 받은 후 언어모델, 발음 사전 및 단어를 구성하는 음소 HMM 구성을 사용하지 않고 단순히 문자를 출력해내는 모델을 시도하였다. 인코더-디코더 구조의 end-to-end 모델에서 음향모델은 인코더에서 담당하고 언어모델은 디코더에서 학습할 수 있도록 하였다. 기본적인 sequence-to-sequence 프레임워크에 attention 메커니즘 (Bahdanau et al., 2015; Chorowski et al., 2014; Chorowski et al., 2015)을 도입하여 인코더의 음성 정보와 디코더의 언어 정보 사이의 상관관계를 학습하게 된다.

LAS 모델은 큰 범주에서 인코더-디코더 구조의 모델이다. 음성 신호로부터 추출한 필터뱅크 기반 스펙트럼 특징을 x, 해당 음성의 정답 문자열을 y라고 하면, 다음 식과 같이 나타낼 수 있다.

$$x = (x_{1,\ldots,} x_t) \tag{1}$$

$$y = (< sos >, y_{1,} \ldots, y_S < eos >) \tag{2}$$

$$y_i \in a,\ldots,z,0,\ldots,9,(space),(,),(.),('),(unknown) \tag{3}$$

여기에서 x는 입력 음성의 특징 벡터열이며, y는 텍스트에서 발견되는 전체 알파벳, 숫자, 공백, 문장 부호, SOS(start-of-

sentence) 토큰, EOS(end-of-sentence) 토큰, unknown 토큰 등으로 이루어져 있다. LAS 모델은 이렇게 정의된 x와 y를 이용하여 인코더 입력에 x, 디코더 입력에 y를 사용한다.

인코더는 Listener라는 이름을 갖는 RNN(recurrent neural network) (Mikolov et al., 2011) 모델을 사용하며, 디코더로 Speller라는 이름을 갖는 RNN 모델을 사용한다. Listener는 피라미드 형식으로 구성된 BLSTM(bidirectional long short-term memory) (Schuster & Paliwal, 1997) 인코더이며, 입력 음성의 특징 벡터열인 x로부터 음향적 잠재 변별 정보를 추출한다. 각 방향마다 256 크기를 갖는 512 크기의 양방향 pBLSTM(pyramid BLSTM)을 3개 쌓아 구성하였는데, 이를 통해 계산량을 2^3만큼 줄일 수 있도록 하였다. Listener의 입력으로는 40차원의 로그 필터뱅크 에너지가 사용되었다.

Speller는 입력 음성 x의 인코더 출력에 대한 attention 정보와 디코더 내부의 상태 정보로부터 맥락 벡터를 생성하고, 이를 기반으로 입력 음성에 해당하는 텍스트 y를 출력한다. 2개 계층의 512 크기를 가지는 LSTM 구조를 사용하였고, 가중치는 −0.1에서 +0.1 사이의 균일 분포에 따라 초기화되었다. 이 과정에서 SOS 토큰과 EOS 토큰을 사용하여 문장의 시작과 끝을 학습하도록 하였다. 32의 빔 크기를 갖는 빔 탐색(beam search) 기법을 통하여 N-best 후보를 생성하는 디코딩을 진행하였다.

LAS 모델은 기존 CLDNN-HMM(CNN, LSTM, and DNN-HMM) (Sainath et al., 2015) 모델이 단어오류율 8%의 성능을 보이는 평가셋에 대해 단어오류율 14.1% 의 성능을 보였다. 이 성능은 단어 발음 사전 또는 별도의 언어모델을 사용하지 않은 것이었으며, 추가적으로 외부 언어모델을 사용하는 경우 단어오류율은 10.3%로 개선되었다. 잡음이 있는 조건에 대해서는 CLDNN–HMM의 단어오류율이 8.9%이었고, LAS 모델은 단어오류율 16.5%, 외부 언어모델을 사용하는 경우 12.0%의 단어오류율을 보였다.

2.2. ESPnet: End-to-end speech processing toolkit

ESPnet(Watanabe et al., 2018)은 end-to-end 음성 인식에 초점을 둔 시스템이다. 딥러닝 모델에 기반하며, Kaldi(Povey et al., 2011) 오픈 소스 음성 인식 툴킷의 데이터 처리 기능을 통해 다양한 음성 인식 실험이 가능한 오픈 소스이다. ESPnet은 hybrid CTC(connectionist temporal classification)-attention 기반 모델(Graves et al., 2006) 및 end-to-end 기반 모델 등의 구조를 가지고 있지만, 여기서는 인코더-디코더 형태를 갖는 end-to-end 모델을 비교 소개한다.

ESPnet의 인코더는 특정 길이의 음성 특징 벡터열이 입력되면 서브 샘플링 기능을 갖는 pBLSTM를 통해 음향적 잠재 정보를 추출하는 방식을 따르는데, 입력 음성에 대한 특징 추출로부터 얻은 멜 필터뱅크 에너지 파라미터에 VGG16(Visual Geometry Group, 16 layers) (Simonyan & Zisserman, 2014) 구조를 적용하여 local 정보를 학습한 후, VGG16 출력에 순차적 정보를 학습하는 BLSTM을 추가하고, 마지막으로 attention score를 계산하는 방식으로 구현되었다.

디코더는 RNN 모델로 구성되어 있는데, 인식 대상 어휘로 SOS 토큰과 EOS 토큰을 추가하고 텍스트 데이터에 단어 임베딩(Levy & Goldberg, 2014)을 적용하여 얻은 단어 벡터를 이용하여 학습하게 된다.

인코더와 디코더 사이에 attention 메커니즘을 사용하였는데, location-aware attention 메커니즘(Chorowski et al., 2015)과 dot-product attention 메커니즘(Luong et al., 2015) 모두를 제공하나, 전자의 방법이 음성 인식 성능에 유리하였고, 후자의 방법은 빠른 계산을 요구하는 경우에 유리하였다.

ESPnet은 Wall Street Journal(WSJ) 데이터베이스(Paul & Baker, 1992)의 eval92 데이터셋을 기준으로 7.6%의 음절오류율을 달성하였다. BLSTM을 6개로 증가한 경우 5.9%의 음절오류율을 보였고, 외부 언어모델을 사용한 경우 음절오류율을 3.8%까지 개선하였다.

3. 제안된 End-to-End 음성 인식 모델

3.1. 모델 구조

본 논문에서 제안하는 end-to-end 모델 구조는 CNN-LSTM 형태의 인코더를 사용한다. 음향모델을 구축하는 인코더에서는 CNN 기반 VGG16 구조를 통하여 입력 특징 벡터의 local 정보를 추출하고, 음성 데이터의 순차적인 정보를 추출하기 위해 VGG16에서 추출된 local 정보를 이용하여 BLSTM을 적층함으로써 최종적으로 음향적 잠재 변별 정보를 학습하게 된다.

제안된 구조에서 VGG16 부분은 conv-16 2번에 maxpooling (Ranzato et al., 2007)을 적용하고, 다시 conv-32 3번에 maxpooling 적용 후 conv-64 2번에 maxpooling을 하는 방식으로 구성되었다. 이후에 batch normalization(Ioffe & Szegedy, 2015)과 ReLU (rectified linear units, Nair & Hinton, 2010), dropout(Srivastava et al., 2014)을 적용하였으며, dropout의 계수는 0.3으로 하였다. 여기서 얻어진 VGG16 출력을 BLSTM 입력으로 이용하여 시간상의 순차 정보를 표현할 수 있도록 학습하였다.

또한, 인코더에서 다양한 음향적 특성을 표현하기 위해 스펙트럼 데이터 증강 기법을 제안한다. 그림 1에 나타낸 음향적 데이터 증강(acoustic data augmentation, ADA) 방법은 멜 필터뱅크 에너지 열에 대해 매 프레임마다 임의로 특정 필터뱅크의 값을 0으로 masking하여 학습에 활용할 수 있는 데이터를 확보하는 것으로, 프레임 스텝 별로 10% masking 및 20% masking 적용하여 음성 데이터를 3배로 증가하는 효과를 얻었다. 이 방법은 제한된 음성 데이터베이스를 이용하여 음향모델을 효과적으로 학습할 수 있을 것으로 기대된다.

제안된 방법의 디코더는 self-attention 메커니즘으로 sequence 데이터를 효과적으로 학습해내는 Transformer 디코더 구조에 기반하고 있다. 디코더에 입력되는 텍스트의 앞뒤로 SOS 토큰과 EOS 토큰을 추가하였고, hidden layer에 해당하는 벡터의 크기로 단어 임베딩을 수행하였다.

다음으로 음성 입력에 대한 출력 문자열의 예측을 학습하기 위해 scaled dot-product attention 메커니즘과 masked multi head attention(MHA) 메커니즘을 사용하였다. Dot-product attention 메커니즘은 입력 음성의 잠재 변별특성을 학습하는 인코더의 출력과 디코더 입력 텍스트 사이의 관계를 모델링하는 것이고, masked MHA 메커니즘은 입력 음성에 해당하는 단어열을 하나씩 순서대로 디코딩하여 출력하는 음성 인식 과정을 학습하기 위해 디코더 입력 텍스트에서 특정 단어가 이전 단어열에 대해서만 attention이 일어나도록 하고 이후 단어열을 학습하지 못하도록 하는 것이다. 이 과정을 통해 디코더는 현재까지 인식된 단어열에 대해 입력 음성에 적합한 다음 단어를 출력하는 모델을 학습할 수 있게 된다. Masked MHA 메커니즘은 head 개수에 해당하는 서로 다른 필터를 적용하는 효과를 줄 수 있어 주어진 텍스트에 대해 현재 단어와 다른 단어 사이의 관련성을 효과적으로 모델링할 수 있다.

하지만 기본 Transformer 디코더는 입력 음성에 해당하는 텍스트를 디코더 입력으로 하여 인식 결과로 출력하도록 학습하는 방법으로, 학습 데이터에 나타난 언어 표현만을 학습하게 되는 문제가 있고, 이로 인해 제한된 학습 데이터의 경우 다양한 언어모델에 대한 학습이 어려워지게 된다. 제안된 방법에서는 이 문제를 해결하기 위해 기본 Transformer 디코더에 masked 언어정보 예측(masked linguistic prediction, MLP) 기능을 그림 1에서처럼 추가하여 입력 텍스트에 대해 임의의 masking을 하고 masking된 텍스트로부터 원본 텍스트를 예측하도록 디코더를 개선하였다. 따라서 언어정보 표현을 위한 모델을 강화할 수 있을 것으로 기대된다.

제안된 end-to-end 음성 인식 모델은 앞에서 기술된 인코더와 디코더의 결합 구조로 되었으며, 인코더에는 순차적인 음향 정보를 표현하고, 디코더는 MLP를 통해 self-attention 기반 다양한 언어 표현을 학습해내는 sequence-to-transformer(SEQFORMER) 구조를 가지게 된다.

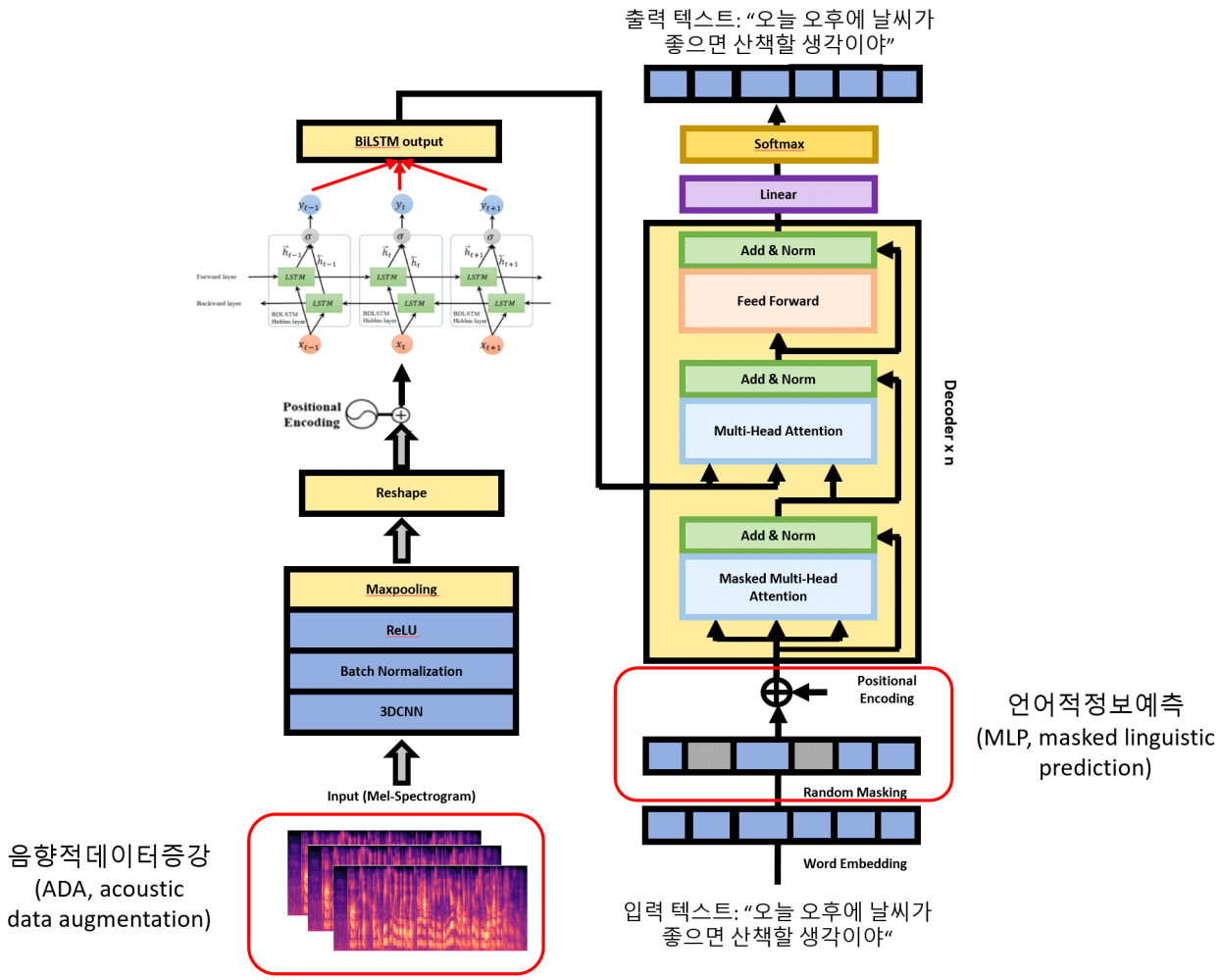

그림 1. 제안된 end-to-end 음성 인식 구조
Figure 1. The proposed end-to-end speech recognition architecture

3.2. 상세 모델변수 설정

본 논문에서 hidden layer 벡터 차원은 512로 사용하였으며, CNN은 3D-CNN을 사용하였다. 인코더의 BLSTM 계층은 3개를 사용하였다. 디코더에서 반복되는 MHA 구조는 6개 계층으로 구현되었고, MHA의 head 개수는 8을 사용하였다. 디코더에 사용된 dropout 계수는 0.1이었다. 전체 모델을 학습할 때 초기 학습률은 0.0001이었고, decay step 4,000 및 decay rate 0.96으로 학습률을 조정하였다. 또한, 학습에 사용된 배치 사이즈는 32로 하였다.

4. 음성 인식 실험

제안된 방법의 성능이 기존 방법 대비 제한된 데이터 조건에서 우수한 성능을 가짐을 보여주기 위한 음성 인식 실험을 진행한다. 어린이 및 노인 음성에 대해 제한된 데이터 환경의 인식 성능 평가를 진행하였다. 먼저 사용한 어린이 및 노인 음성 데이터베이스에 대해 소개하고, 음성 인식 평가 과정 및 결과를 통해 제안된 방법과 기존 방법을 비교한다.

4.1. 어린이 음성 데이터베이스

본 논문에서는 2개의 공개된 어린이 음성 데이터셋을 이용하여 음성 인식 실험을 진행하였다.

4.1.1. 한국전자통신연구원(ETRI) 어린이 음성 데이터셋

ETRI에서는 어린이 음성 데이터셋을 구축하여 공개하고 있다.[1] 본 데이터셋은 '음성인터페이스 개발을 위한 어린이 음성 데이터'로서 어린이 음향 모델 훈련용이며, 초등학교 1–4학년을 대상으로 녹음을 진행하였다. 아이폰5, 갤럭시S4 및 마이크로폰을 이용하여 동시에 16 kHz로 녹음하였다. 데이터셋의 구성은 총 50명이 100개의 발화를 한 것으로, 9.68시간으로 이루어져 있다.

음성 인식 실험을 위해 총 16,200개의 데이터 가운데 학습용 데이터로 12,960개, 검증용 데이터로 1,620개, 그리고 테스트 데이터로 1,620개를 임의로 나누어 진행하였다.

4.1.2. 엘솔루 어린이 음성 데이터셋

엘솔루에서는 'KETI 지능정보 플래그쉽 R&D 어린이 음성 데이터셋'을 구축하여 공개하고 있다.[2] 본 데이터베이스는 초등학교 1–6학년을 대상으로 조용한 환경에서 스마트폰으로 녹음을 진행하였다. 300명이 각각 60–100문장을 발화하였으며, 전체 음성 데이터 개수는 25,369개이고, 전체 시간은 22.4시간이다.

실험을 위해서는 총 25,369개의 음성 데이터 가운데 학습용 데이터로 20,295개, 검증용 데이터로 2,537개, 그리고 테스트 데이터로 2,537개를 임의로 나누어 진행하였다.

4.2. 노인 음성 데이터베이스

노인 음성 데이터베이스는 마인즈랩과 ETRI에서 공개한 노인 음성 데이터셋을 사용하였다.[3] 마인즈랩과 ETRI에서 구축한 노인 음성 데이터셋인 'VOTE400'은 고령자와 로봇 간 원활한 음성 교류를 가능하게 하는 음성 인식 시스템을 개발하기 위하여 수집한 대용량 한국어 노인 음성 데이터이다. 발화자는 다양한 지역별로 분포되어 있으며, 총 300시간으로 구성되어 있다. 음성 데이터는 16 bit MONO PCM 포맷으로 44,100 Hz, 11,050 Hz 등 다양한 sampling rate로 저장되어 있다. 대화체 및 낭독체 음성 데이터셋으로 구성되어 있는데, 대화체 음성 데이터는 두 사람이 연속적으로 대화한 것을 녹음한 음성이며, 낭독체 음성 데이터는 개별 문장에 해당하는 음성이 녹음되어 있다. 본 논문에서는 낭독체 음성 데이터셋의 약 100시간에 대해서만 실험하였으며, 전체 데이터 111,814개 가운데 학습 데이터로 89,426개, 검증용 데이터로 11,178개, 그리고 테스트 데이터로 11,179개를 임의로 나누어 제안된 방법의 평가를 수행하였다.

4.3. 실험 결과 및 분석

음성 인식 실험에서는 음절인식률을 이용하여 성능을 평가하였다. 어린이 음성 데이터셋과 노인 음성 데이터셋에 대한 성능 평가 결과는 다음과 같다.

4.3.1. 어린이 음성 인식 실험 결과

어린이 음성 데이터셋에 대한 음성 인식 실험 결과는 표 1과 같다. 엘솔루 어린이 데이터셋에 대해 학습 및 평가한 경우 LAS 모델은 89.49%의 음절인식률을 얻었다. 제안된 방법의 기본 구조인 SEQFORMER 방법은 88.53%의 음절인식률을 얻었다. 기본 SEQFORMER 구조에 음향적 데이터 증강 방법인 ADA를 적용하여 학습하고 평가한 경우 음절인식률은 94.5%이다. 다양한 언어 표현을 예측하는 MLP 방법을 적용하는 경우 95.34%의 음절인식률을 얻었으며, 제안한 2가지 방법을 모두 적용하면 96.85%까지 음절인식률이 개선됨을 알 수 있었다. 이때 MLP 방법에 적용된 임의의 masking 비율은 20%이었다.

엘솔루 어린이 음성 데이터셋 보다 상대적으로 개수가 적은 ETRI 어린이 음성 데이터셋으로 학습, 평가하는 경우 LAS 모델은 99.12%의 음절인식률을 얻었다. 제안된 모델의 경우 기본 SEQFORMER 구조에 대해 96.43%의 음절인식률을 보였다, 데이터의 다양성을 얻기 위해 제안한 방법을 적용하는 경우 ADA 방법은 97.43%의 음절인식률을 보였고, 다양한 언어표현을 예측하는 MLP 방법은 97.79%의 음절인식률을 얻었다. 그리고 두

---

1 http://etri.re.kr/aiopen
2 http://www.aihub.or.kr/keti_data_ board/language_intelligence
3 https://ai4robot.github.io/mindslab-etri-vote400/#

가지 방법을 모두 적용하는 경우 98.32%의 음절 인식률을 얻었다. 이것은 데이터셋의 규모가 매우 적은 상황으로 LAS와 제안된 방법의 성능 차이는 일반화하기는 어려운 점이 있다. 따라서 두 평가셋에 대한 평균 성능을 측정하였으며, 이 경우 LAS 방법이 93.25%의 음절인식률을 가지는데 비해 제안된 방법은 ADA 방법과 MLP 방법을 모두 적용한 구조에 있어 97.41%의 음절인식률을 보였다. 이 결과로부터 제안된 방법이 기존 end-to-end 음성 인식 모델에 비해 제한된 데이터 환경에서 우수한 성능을 보임을 알 수 있다.

표 1. 어린이 음성 인식 결과
Table 1. Children speech recognition results

모델		엘솔루 어린이 데이터셋 음절인식률 (%)	ETRI 어린이 데이터셋 음절인식률 (%)	평균 음절 인식률 (%)
LAS		89.49	99.12	93.25
제안된 방법	Basic SEQFORMER	88.53	96.43	91.61
	SEQFORMER +ADA	94.50	97.43	95.64
	SEQFORMER +MLP	95.34	97.79	96.30
	SEQFORMER +ADA+MLP	96.83	98.32	97.41

LAS, listen, attend, and spell; SEQFORMER, sequence-to-transformer; ADA, acoustic data augmentation; MLP, masked linguistic prediction.

어린이 음성 데이터셋 실험의 경우, 두 데이터셋의 학습 데이터가 적은 상황이었다. 이 문제를 인코더의 음향적 증강 및 디코더의 언어적 변이 예측을 통해 인식 성능 개선을 확인할 수 있었으나, 일정 규모로 이루어진 제한된 데이터 환경에서 제안된 방법의 유용함을 평가할 필요가 있다. 이를 위해 일정 규모의 데이터로 구성된 노인 음성 인식 실험을 다음과 같이 진행하였다.

4.3.2. 노인 음성 인식 실험 결과

노인 음성 인식 평가 환경도 제한된 학습 데이터 조건이지만, 어린이 음성 데이터베이스에 비해 상대적으로 용량이 크므로 제안된 방법을 평가하기에 적절하다고 볼 수 있다. 제한된 데이터를 이용하여 학습한 성능과 데이터를 추가하여 학습한 성능을 비교하여 제안된 방법이 제한된 데이터 조건에서 모델 학습을 강인하게 함을 평가하였다. 이를 위해 89,426개의 학습 데이터 가운데 80%인 71,540개로 제한된 데이터를 대상으로 학습을 진행하였고, 검증 및 테스트 데이터는 4.2 절에 기술한 대로 사용하였다. 이에 대한 실험 결과를 표 2에서 확인할 수 있다.

먼저 전체 학습 데이터의 80%를 이용하여 학습 및 평가한 경우, LAS 모델은 95.18%의 음절 인식률을 얻었다. 제안된 모델의 경우 기본 SEQFORMER 구조에 대해 96.21%의 음절인식률을 보였다. 데이터의 다양성을 얻기 위한 ADA 방법은 96.66%의 음절인식률을 보였고, 다양한 언어표현을 예측하는 MLP 방법은 97.61%의 음절인식률을 얻었다. 그리고 두 가지 방법을 모두 적용하는 경우 97.69%의 음절 인식률을 얻었다.

제안된 방법에서 기본 SEQFORMER 구조에 비해 ADA 및 MLP 구조를 적용하는 경우 약 39.1%의 오류감소율을 얻었으며, 이를 통해 제안된 방법이 음향 정보를 풍부하게 학습하는 동시에 언어모델을 강력하게 만드는 것을 알 수 있었다.

표 2. 제한된 학습 데이터에 대한 노인 음성 인식 결과
Table 2. Elderly speech recognition results for limited training data

모델		노인 데이터셋 음절인식률(%)
LAS		95.18
제안된 방법	Basic SEQFORMER	96.21
	SEQFORMER +ADA	96.66
	SEQFORMER +MLP	97.61
	SEQFORMER +ADA+MLP	97.69

LAS, listen, attend, and spell; SEQFORMER, sequence-to-transformer; ADA, acoustic data augmentation; MLP, masked linguistic prediction.

표 3은 89,426개의 전체 학습 데이터를 이용하여 학습한 모델을 평가한 결과이다. 기존의 LAS 모델로부터 97.31%의 음절 인식률을 얻었다. 제안된 기본 SEQFORMER 구조는 96.44%의 음절인식률을 얻었고, ADA를 적용한 경우 음절인식률 96.70%, MLP를 적용한 경우 음절인식률 97.66%를 얻었다. 또한 ADA와 MLP를 모두 적용한 방법은 97.75%의 음절인식률을 얻었다.

표 3. 학습 데이터 증가에 대한 노인 음성 인식 결과
Table 3. Elderly speech recognition results for increased training data

모델		노인 데이터셋 음절인식률(%)	학습 데이터 증가에 따른 성능 차이(%)
LAS		97.31	+2.18
제안된 방법	Basic SEQFORMER	96.44	+0.23
	SEQFORMER +ADA	96.70	+0.04
	SEQFORMER +MLP	97.66	+0.05
	SEQFORMER +ADA+MLP	97.75	+0.06

LAS, listen, attend, and spell; SEQFORMER, sequence-to-transformer; ADA, acoustic data augmentation; MLP, masked linguistic prediction.

위의 결과를 통해 LAS 모델은 학습 데이터 개수가 증가함에 따라 성능 개선 정도가 높음을 알 수 있다. 이것은 기존의 end-to-end 음성 인식 방법이 많은 학습 데이터를 요구함을 의미하며, 노인 음성 인식과 같이 제한된 데이터 조건에는 유용하지 않음을 보여준다. 반면, 제안된 방법의 경우 학습 데이터 80%를 사용한 모델과 데이터를 추가하여 100%를 사용한 모델 사이에 성능 차이가 크지 않음을 확인할 수 있다. 이것은 제안된 방법

이 제한된 학습 데이터 환경에 효율적으로 강인한 모델 학습을 할 수 있음을 의미한다. 따라서 음향적 다양성과 언어 표현의 다양성을 제한된 학습 데이터로부터 모델링하도록 제안된 방법이 강인한 end-to-end 음성 인식 엔진을 구현하는데 활용할 수 있음을 알 수 있다.

5. 결론

현재 개발되고 있는 음성 인식 시스템은 대부분 성인 남녀를 기준으로 음성 인식을 수행하도록 구현되어 있다. 음성 인식 모델의 학습을 위해 성인 남녀를 대상으로 음성 데이터베이스를 구축하는 것이 유리하기 때문이다. 따라서, 노인과 어린이 등의 음성 인식을 위해서는 제한된 음성 데이터베이스의 문제를 해결해야 한다.

본 논문에서는 이를 위해 제한된 데이터를 이용하여 강인한 학습이 가능하도록 음향적 정보를 증강하는 재귀적 뉴럴 네트워크 기반의 인코더와 언어적 변이의 다양성을 학습하도록 개선된 Transformer 디코더로 이루어진 SEQFORMER 구조의 end-to-end 방식을 제안한다. 제한된 학습 데이터베이스를 이용하는 성능 평가에 있어, 데이터가 극히 제한된 어린이 음성 인식 실험에서는 일반화가 어려웠던 반면 노인 음성 인식 실험에서는 음향적 다양성과 언어 표현의 다양성을 학습하여 강인한 end-to-end 음성 인식 엔진을 구현하는데 제안된 방법이 활용될 수 있음을 보였다.

References

Bahdanau, D., Cho, K., & Bengio, Y. (2015, January). Neural machine translation by jointly learning to align and translate. *Proceedings of the 3rd International Conference on Learning Representations*, *ICLR 2015*. San Diego, CA.

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.

Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., & Bengio, Y. (2015). Attention-based models for speech recognition. *Advances in Neural Information Processing Systems*, *2015*, 577-585.

Chorowski, J., Bahdanau, D., Cho, K., & Bengio, Y. (2014, December). End-to-end continuous speech recognition using attention-based recurrent NN: First results. *Proceedings of the NIPS 2014 Workshop on Deep Learning*. Montreal, Canada.

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.

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.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Proceedings of the 32nd International Conference on Machine Learning*, *PMLR* (pp. 448-456). Mountain View, CA.

Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In Z. Ghahramani., M. Welling, C. Cortes, N. Lawrence & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (pp. 2177-2185). Red Hook, NY: Curran Associates.

Luong, M. T., Pham, H., & Manning, C. D. (2015, September). Effective approaches to attention-based neural machine translation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 1412-1421). Lisbon, Portugal.

Mikolov, T., Kombrink, S., Burget, L., Černocký, J., & Khudanpur, S. (2011, May). Extensions of recurrent neural network language model. *Proceedings of the 2011 IEEE International Conference on Acoustics*, *Speech and Signal Processing (ICASSP)* (pp. 5528-5531). Prague, Czech Republic.

Nair, V., & Hinton, G. E. (2010, January). Rectified linear units improve restricted boltzmann machines. *Proceedings of the 27th International Conference on International Conference on Machine Learning (ICML)* (pp. 807-814). Haifa, Israel.

Paul, D. B., & Baker, J. (1992, February). The design for the wall street journal-based CSR corpus. *Proceedings of the Speech and Natural Language: Proceedings of a Workshop Held at Harriman.* New York, NY.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., ... Silovsky, J. (2011, December). The Kaldi speech recognition toolkit. *Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. Big Island, HI.

Ranzato, M. A., Huang, F. J., Boureau, Y. L., & LeCun, Y. (2007, June). Unsupervised learning of invariant feature hierarchies with applications to object recognition. *Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1-8). Minneapolis, MN.

Sainath, T. N., Vinyals, O., Senior, A., & Sak, H. (2015, April). Convolutional, long short-term memory, fully connected deep neural networks. *Proceedings of the 2015 IEEE International Conference on Acoustics*, *Speech and Signal Processing (ICASSP)* (pp. 4580-4584). Brisbane, Australia.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11),

2673-2681.
Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv*. Retrieved from https://arxiv.org/abs/1409.1556
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929-1958.
Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Z. Ghahramani., M. Welling, C. Cortes, N. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (pp. 3104-3112). Red Hook, NY: Curran Associates.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In U. von Luxburg, A. Bengio, R. Fergus, R. Garnett, I. Guyon, H. Wallach, & S. V. N. Vishwanathan (Eds.), *Advances in neural information processing systems* (pp. 5998-6008). Red Hook, NY: Curran Associates.
Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Soplin, N. E. Y., ... Renduchintala, A. (2018). Espnet: End-to-end speech processing toolkit. *arXiv*. Retrieved from https://arxiv.org/abs/1804.00015

• **김준우 (June-Woo Kim)**
경북대학교 인공지능학과 석사과정
대구광역시 북구 대학로 80 경북대학교
Tel: 053-940-8616
Email: kaen2891@gmail.com
관심분야: 음성 인식, 음성 변환, 음성 합성, 딥러닝

• **정호영 (Ho-Young Jung)** 교신저자
경북대학교 인공지능학과 교수
대구광역시 북구 대학로 80 경북대학교
Tel: 053-950-2337
Email: hoyjung@knu.ac.kr
관심분야: 음성 인식, 음성 변환, 자연어 이해, 딥러닝

제한된 학습 데이터를 사용하는 End-to-End 음성 인식 모델*

김 준 우 · 정 호 영

경북대학교 인공지능학과

국문초록

음성 인식은 딥러닝 및 머신러닝 분야에서 활발히 상용화 되고 있는 분야 중 하나이다. 그러나, 현재 개발되고 있는 음성 인식 시스템은 대부분 성인 남녀를 대상으로 인식이 잘 되는 실정이다. 이것은 음성 인식 모델이 대부분 성인 남녀 음성 데이터베이스를 학습하여 구축된 모델이기 때문이다. 따라서, 노인, 어린이 및 사투리를 갖는 화자의 음성을 인식하는데 문제를 일으키는 경향이 있다. 노인과 어린이의 음성을 잘 인식하기 위해서는 빅데이터를 구축하는 방법과 성인 대상 음성 인식 엔진을 노인 및 어린이 데이터로 적응하는 방법 등이 있을 수 있지만, 본 논문에서는 음향적 데이터 증강에 기반한 재귀적 인코더와 언어적 예측이 가능한 transformer 디코더로 구성된 새로운 end-to-end 모델을 제안한다. 제한된 데이터셋으로 구성된 한국어 노인 및 어린이 음성 인식을 통해 제안된 방법의 성능을 평가한다.

핵심어: 음성 인식, 종단간 음성 인식, 적은 데이터 음성 인식

* 본 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (2020-0-01808, 복합정보 기반 예측 지능 혁신 기술 연구).