The Korean Journal of Applied Statistics
2023, Vol. 36, No. 1, 33–48

DOI: https://doi.org/10.5351/KJAS.2023.36.1.033
Print ISSN1225-066X / Online ISSN 2383-5818

Deep learning-based speech recognition for Korean elderly speech data including dementia patients

Jeonghyeon Mun*[a], Joonseo Kang*[a], Kiwoong Kim[bcd], Jongbin Bae[bc], Hyeonjun Lee[e], Changwon Lim[1,a]

[a]Department of Applied Statistics, Chung-Ang University;
[b]Department of Neuropsychiatry, Seoul National University Bundang Hospital;
[c]Department of Psychiatry, Seoul National University;
[d]Department of Brain and Cognitive Sciences, Seoul National University; [e]Sevenpointone

Abstract

In this paper we consider automatic speech recognition (ASR) for Korean speech data in which elderly persons randomly speak a sequence of words such as animals and vegetables for one minute. Most of the speakers are over 60 years old and some of them are dementia patients. The goal is to compare deep-learning based ASR models for such data and to find models with good performance. ASR is a technology that can recognize spoken words and convert them into written text by computers. Recently, many deep-learning models with good performance have been developed for ASR. Training data for such models are mostly composed of the form of sentences. Furthermore, the speakers in the data should be able to pronounce accurately in most cases. However, in our data, most of the speakers are over the age of 60 and often have incorrect pronunciation. Also, it is Korean speech data in which speakers randomly say series of words, not sentences, for one minute. Therefore, pre-trained models based on typical training data may not be suitable for our data, and hence we train deep-learning based ASR models from scratch using our data. We also apply some data augmentation methods due to small data size.

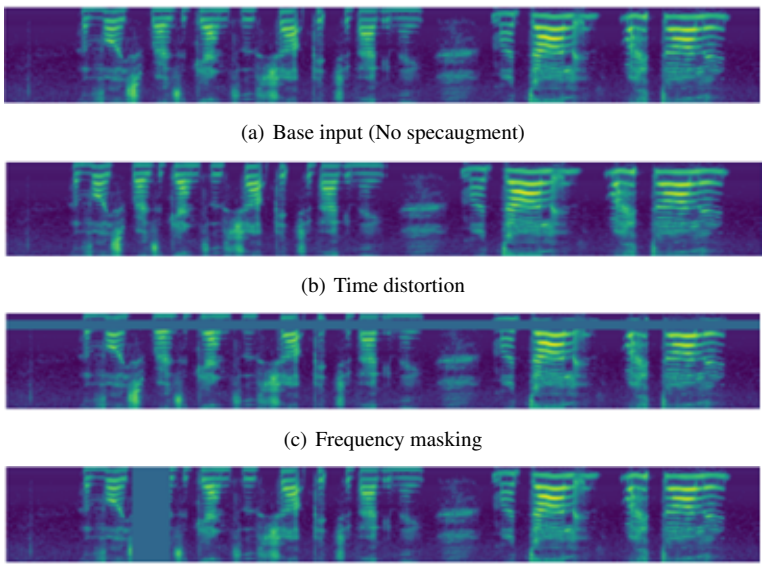
Keywords: Korean elderly speech data, automatic speech recognition, deep-learning, data augmentation

1. 서론

딥러닝(deep-learning) 기술의 도래는 다양한 인공지능 분야에서 기존의 한계를 뛰어넘어 급격한 성장을 이루고 있으며 그 성장은 자동 음성 인식(automatic speech recognition; ASR) 분야에서도 마찬가지이다 (Malik 등, 2021). 자동 음성 인식은 음성 신호를 인식하고 텍스트로 변환하는 기술이다. 자동 음성 인식 시스템 구축은 일반적으로 음성으로부터 feature를 뽑아내는 것으로 시작한다. Mel-frequency cepstral coefficients (MFCC)와 filter bank, Mel-spectrogram과 같은 방법들이 feature 추출에 사용된다 (Chakraborty 등, 2014). 이를 바탕으로 자동 음성 인식 모델은 스마트 스피커 등 다양한 IoT 디바이스와 챗봇 등의 서비스에 활용되고 있으며, 인공지능 비서, 동시통역, 인공지능 튜터 등 다양한 애플리케이션에 활용되고 있다.

This research was supported by the Chung-Ang University Graduate Research Scholarship in 2021.
This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science, ICT (NRF-2021R1F1A1056516).
*These authors are co-first authors.
[1] Corresponding author: Department of Applied Statistics, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Korea. E-mail: clim@cau.ac.kr

Published 28 February 2023 / journal homepage: http://kjas.or.kr



(a) Base input (No specaugment)

(b) Time distortion

(c) Frequency masking

(d) Time masking

Figure 1: *SpecAugment example (Park et al., 2019).*

현재 음성인식 연구의 한계점 중 하나는 음성 데이터가 대부분 서양권 언어, 특히 영어에 치중된다는 점이다. 대부분의 모델들은 하나의 문장 음성을 input으로 받아 학습하여 26자의 alphabet 중 하나를 연속하여 출력하는 구조로 되어 있다. 따라서 대상 언어가 알파벳을 사용하지 않는 언어일 경우 모델에 따라서 큰 폭으로 성능이 떨어지기도 한다. 이는 동아시아권 언어, 즉 한국어, 일본어, 중국어 등에서 두드러지게 나타나며, output으로 나올 수 있는 글자 후보군의 가짓수가 alphabet에 비해 매우 많기 때문이다. 한글로 표기할 수 있는 문자는 10,000여개 가까이 존재하며, 따라서 data crawling 등을 통해 자주 사용되는 문자만을 output으로 나오도록 제한하는데, 이를 고려해도 한국어는 약 2,000자가 넘는 단어 개수가 후보군으로 선정될 수 있다 (Kim과 Lee, 2020).

또한, 국내외 서비스 중인 자동 음성 인식 시스템 대부분은 성인 남녀의 음성을 잘 인식하는 편이다. 이는 자동 음성 인식 모델의 일반적인 학습 데이터는 발음과 발성이 명확하고 이상적인 환경에서 녹음된 음성들로 구성되어 있기 때문이다. 현재 음성 인식에서 자주 사용되는 오픈 데이터셋으로 LibriSpeech (Panayotov 등, 2015) 데이터셋이 있다. 이 데이터셋은 2,484명의 화자가 약 1,000시간가량의 오디오 북을 녹음한 음성 데이터이며, 주로 문장으로 이루어져 있으며, 발음이 정확한 20~50대의 성인 발화자들로 이루어져 있다. 국내에서 연구 목적으로 제공하는 AI Hub의 음성 데이터셋도 대개 성인 남녀 기준으로 구성되어 있으며, 음성 인식 서비스를 제공하는 기업에서도 데이터 수집의 어려움과 비용 등의 문제로 성인 남녀의 음성을 중심으로 엔진을 개발하고 있는 실정이다.

본 논문에서는 한국어 음성 데이터를 다루고 있으며, 음성 데이터의 발화자 대부분은 60세 이상의 고령층이며 일부는 치매환자이다. 또한, 음성 데이터는 발화자들이 1분 동안 동물이나 채소와 같은 일련의 단어를 무작위로 말하는 특징을 가지고 있다. 고령층의 목소리는 젊은 성인과는 다른 특징을 가지고 있다. 혀의 활동 범위 및 두께가 감소하고, 말하는 속도가 느리며, 침묵하는 부분이 증가할뿐더러 발음의 정확성이 떨어지는 것으로 알려져 있다 (Lee와 Gwon, 2014). 따라서 고령층 대상의 음성 인식은 성인 대상의 음성 인식보다 정확도가 높지 않을 것으로 예상된다 (Young과 Mihailidis, 2010). ASR 모델 학습을 위해 설계된 데이터셋에서

Table 1: Values of parameters of SpecAugment policy applied to our data. m_f and m_t denote the number of frequency and time maskings, respectively.

Policy	F	m_f	T	m_t
Value	27	2	20	4

발화자의 발음은 일반적으로 정확하다. 각 데이터 세트마다 다양한 발화자가 존재할 수 있지만, 정상적인 경우 대부분의 발화자가 고령층인 데이터셋은 거의 없다. 더불어, 치매를 앓고 있는 고령층의 음성 데이터는 묵음 구간이 많이 존재하며, 일반 고령층보다 발음이 더 부정확하기 때문에 자동 음성 인식의 정확도가 더 떨어질 것으로 예상된다. 또한, 우리의 데이터는 약 50시간의 양을 녹음한 음성 데이터로 LibriSpeech 데이터셋처럼 크지 않다. 따라서 우리 데이터는 사이즈가 작고 문장이 아닌 단어들이 녹음되어 있으며, 발음이 부정확할 수 있는 발화자들로 이루어져 있는 음성 데이터이기 때문에 일반적인 학습 데이터를 기반으로 하는 사전 학습 모델은 적합하지 않을 수 있다.

본 논문에서는 LibriSpeech 등의 데이터셋을 기반으로 개발된 최신 딥러닝 기반 자동 음성 인식 모델들을 우리 데이터, 치매 환자 포함 고령층 한국어 음성 데이터로 학습시키고 성능을 비교하고자 한다. 먼저, 전처리 단계로 음성 인식에 불필요한 부분인 무음 부분을 데이터에서 제거하여 데이터의 길이를 줄인다. 둘째, 데이터 증강을 적용하여 음성의 속도와 피치를 변경하고 노이즈를 추가한다. 이러한 증강을 통해 데이터 크기가 4배 증가한다. 셋째, 주파수를 mel-scale로 변환하여 음성 특징을 추출하는 mel-spectrogram을 오디오에 적용한다. 그 후, ASR에 또 다른 데이터 증강 방법인 Park 등 (2019)이 제안한 SpecAugment를 적용하여 모델을 학습한다. 영어 데이터의 경우 모델의 성능 평가 지표로 word error rate (WER)를 사용하지만, 한국어의 경우 WER이 아닌 character error rate (CER)를 사용하여 모델을 평가하는 것이 적절하다. 또한, 우리가 사용한 음성인식 모델의 결과와 구글, 네이버 파파고 API를 이용한 음성인식 결과를 비교한다. 이를 통해 우리 데이터셋에 적합한 모델을 찾고자 한다.

본 논문의 주요 공헌은 다음과 같다.

1) 우리는 무음 구간 제거, 데이터 증강 등의 데이터 전처리를 통해 부족한 음성 인식 데이터를 활용하는 작업을 진행했다.
2) 우리는 공개 음성 데이터베이스가 아닌 치매환자를 포함한 대부분의 60세 이상 노인 음성 데이터를 사용하여 여러 음성인식 모델들의 성능 비교를 했다.
3) 여러 모델들과 음성인식 API를 비교하여 우리 데이터에 적합한 모델을 찾았다.

본 논문의 나머지 부분은 다음과 같이 구성되어 있다. 2장에서는 관련 연구들을 설명한다. 3장에서는 우리가 사용하는 데이터셋에 대한 세부 정보 및 전처리와 실험 결과를 제공한다. 마지막으로, 4장에서는 우리의 결론을 도출하고 미래의 연구 방향에 대해 논의한다.

2. 관련 연구법

2.1. SpecAugment

SpecAugment는 Park 등 (2019)이 제안한 ASR의 데이터 증강 방법이다. 기존의 데이터 증강 방법은 오디오가 spectrogram으로 변경되기 이전에 적용하여 모델을 학습한다. 하지만, SpecAugment는 오디오 형태가 아닌 spectrogram 형태를 증강하는 방법이다. 데이터 증강이 input features에 직접적으로 적용되기 때문에, 훈련 속도에 큰 영향을 미치지 않으며, 모델 학습 중 온라인으로 실행될 수 있다. SpecAugment는 모델의 성능을 향상시키고 의도적으로 손상된 데이터를 제공하기 때문에 과적합을 방지할 수 있다.

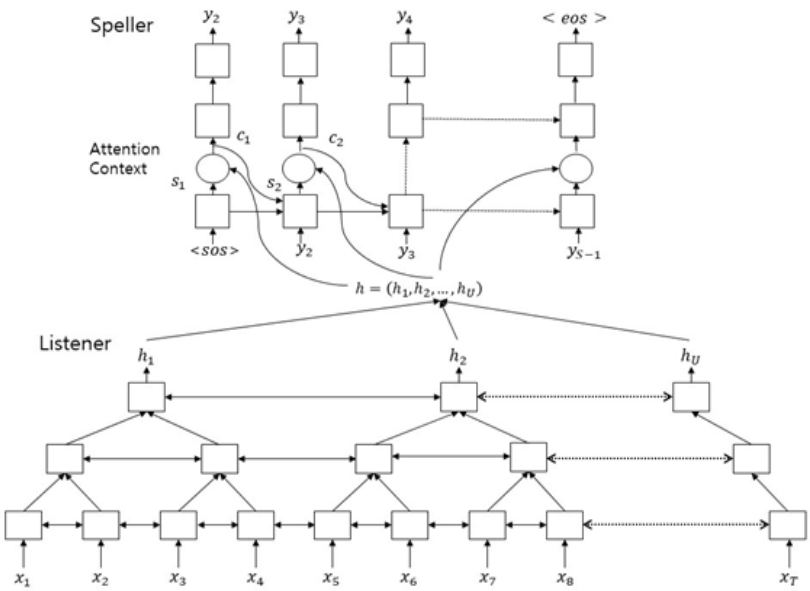


Figure 2: *Architecture of Listen Attend and Spell model. The listener is a pBLSTM encoding input sequence x into high level features h, the speller is an attention-based decoder generating y characters from h (Chan et al., 2016).*

SpecAugment는 3가지의 증대방법이 있으며, time warping, time masking, frequency masking이다. Time warping은 성능에 약간의 영향을 미칠 수 있기 때문에, 우리는 사용하지 않았다. 또한, 전처리 단계에서 음성 속도의 변환을 통해 augmentation을 하기 때문에 time warping을 적용할 필요가 없다. 따라서, 우리는 frequency masking과 time masking을 사용했다. Figure 1은 single input에 SpecAugment가 적용된 예시이다. Figure 1의 (c)는 Frequency masking으로, spectrogram에서 특정 frequency 채널들 $[f_0, f_0 + f)$을 마스킹하는 것이다. f는 0부터 frequency mask parameter F 사이의 균등 분포를 따르는 확률 변수이다. f_0는 $[0, \nu - f)$ 사이의 값으로, ν는 mel frequency 채널들의 수이다. Figure 1의 (d)는 time masking으로 frequency masking과 유사하게 진행된다. Spectrogram에서 특정 time 시점 $[t_0, t_0 + t)$을 마스킹하는 것이다. t는 0부터 time mask parameter T 사이의 균등 분포를 따르는 확률 변수이다. t는 $[0, \tau)$사이의 값으로, τ는 log mel spectrogram의 시간 길이다. Table 1은 우리가 적용한 SpecAugment parameter이다. 우리는 우리 데이터에 frequency masking과 time masking을 적용하였다.

2.2. Listen, Attend and Spell (LAS)

Listen, Attend and Spell (LAS) (Chan 등, 2016) 모델은 입력 음향 신호를 사용하여 한 번에 transcripts를 만드는 음성 인식 모델이다. Figure 2는 LAS 모델의 아키텍처이다.

LAS 모델은 인코더-디코더 형태를 가진다. 인코더는 listener로 불리며 기존의 acoustic model과 유사하다. Listener는 pyramidal bidirectional long short term memory (pBLSTM)로 구성되어 있으며, 입력 음성의 특징 벡터열인 x로부터 정보를 추출한다. Listen 함수는 입력신호 $\boldsymbol{x}$를 high level representation $\boldsymbol{h} = (h_1, \ldots, h_U)$로

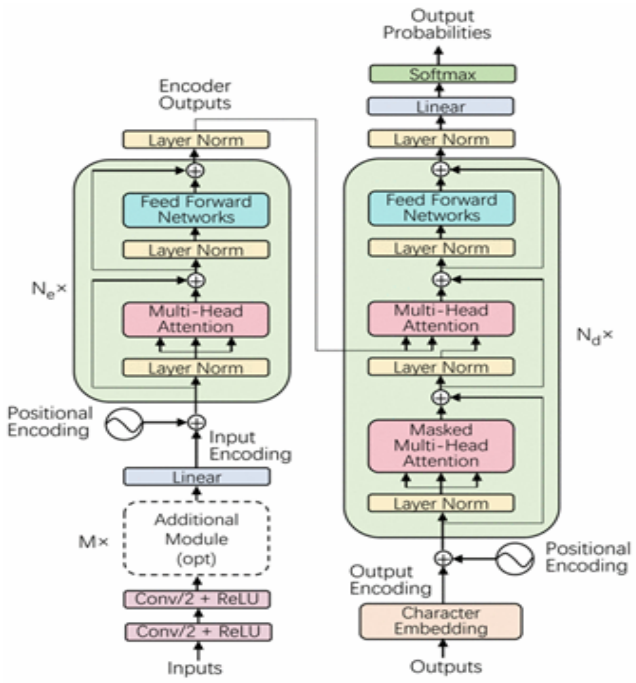


Figure 3: *Architecture of speech-transformer model (Dong et al., 2018).*

변환하고, AttendAndSpell 함수는 high level representation인 $\boldsymbol{h}$로부터 문장의 확률 분포를 만든다.

$$\begin{aligned} \boldsymbol{h} &= \text{Listen}\,(x)\,, \\ P\,(\boldsymbol{y}\mid \boldsymbol{x}) &= \text{AttendAndSpell}\,(h, y)\,. \end{aligned} \tag{2.1}$$

Listen 연산에서는 피라미드 구조의 bidirentional long short term memory (BLSTM)이 사용된다. 이러한 구조를 사용하는 이유는 입력신호 $\boldsymbol{x}$의 길이를 $\boldsymbol{h}$의 길이만큼 줄이기 위함이다. i 번째 시간의 j 번째 층으로부터의 출력은 다음과 같이 이전 step과 이전 layer의 스텝 2개를 입력으로 사용한다. 이것은 attention model에서 짧은 길이로부터 유의미한 정보 추출을 가능케 한다.

$$h_i^j = \text{pBLSTM}\left(h_{i-1}^j, \left[h_{2i}^{j-1}, h_{2i+1}^{j-1}\right]\right). \tag{2.2}$$

입력 음성이 listener를 통과하면, speller라고 불리는 디코더의 입력값이 되는 hidden node vector $\boldsymbol{h}$라는 출력값을 생성한다. Speller는 이전에 본 모든 문자에 대한 다음 문자의 조건부 확률 분포를 생성한다. y_i의 분포는 decoder state s_i와 context vector c_i를 사용하여 softmax 출력값을 가진 multi-layer perceptron (MLP)에 의해 계산된다. Decoder state s_i는 이전 decoder state s_{i-1}, 이전 character y_{i-1}, 이전 context vector c_{i-1}에 의해 생성된다. context vector c_i는 i 번째 시간 단계에서의 decoder state s_i와 다른 u 번째 시간 단계에서의 vector ($h_u \in \boldsymbol{h}$)의 attention mechanism을 통해 생성된다. 이는 다음과 같다.

$$\begin{aligned} s_i &= \text{LSTM}\,(s_{i-1}, y_{i-1}, c_{i-1})\,, \\ c_i &= \text{AttentionContext}\,(s_i, h_u)\,, \\ P\,(y_i \mid \boldsymbol{x}, y_{<i}) &= \text{MLP}\,(s_i, c_i)\,. \end{aligned} \tag{2.3}$$

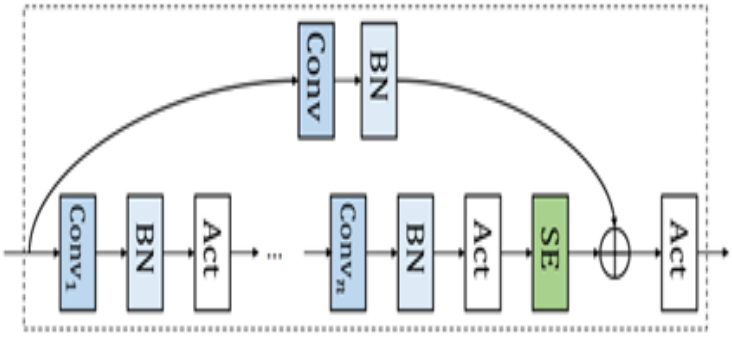


Figure 4: *A convolution block C_i includes many convolutions, each followed by batch normalization and activation. The SE block operates at the output of the last convolution layer (Han et al., 2020).*

2.3. Speech-transformer

Speech-transformer 모델은 음성 feature sequence를 문자 sequence로 변환한다 (Dong 등, 2018). 일반적으로 문자 sequence보다 훨씬 긴 feature sequence는 시간과 주파수를 가진 2차원 spectrogram으로 표현된다. 따라서 Speech-transformer모델은 spectrogram의 구조 locality를 활용하고 데이터의 길이 불일치를 완화해주는 convolutional network을 사용한다. Speech-transformer 또한 인코더와 디코더로 구성된다.

인코더는 Figure 3의 왼쪽과 같다. 먼저, 인코더에서 GPU 메모리의 overflow를 방지하고 문자 길이와 함께 hidden representation을 생성하기 위해 stride 2를 갖는 2개의 3×3 CNN 층을 쌓는다. 다음으로, expressive representations을 추출하기 위해 M개의 추가 모듈을 쌓는다. 이후, 평면 feature map에서 선형 변환을 수행하여 input encoding이라 불리는 d-차원의 벡터들을 얻으며, 모델이 상대적인 위치를 학습할 수 있도록 d-차원의 positional encoding이 추가된다.

$$\mathrm{PE}_{(pos,i)} = \begin{cases} \sin\left(\mathrm{pos}/1000^{\frac{2x}{d_{model}}}\right) & 0 \le i < d_{model}/2 \\ \cos\left(\mathrm{pos}/1000^{\frac{2i}{d_{model}}}\right) & d_{model}/2 \le i < d_{model}, \end{cases} \tag{2.4}$$

여기서 pos는 sequence의 위치를 의미하고, i는 i 번째 차원을 의미한다. PE_{pos+k}는 PE_{pos}의 선형 함수로 나타낼 수 있으며, 임의의 고정 offset k에 의해 positional encoding이 작동한다. 이제 input encoding과 positional encoding의 합을 $\boldsymbol{N}_e$ 인코더 블록의 stack에 입력하여 최종 인코더 출력값을 얻을 수 있으며, 인코더 블록은 2개의 하위 블록이 있다. 첫 번째는 이전 블록 출력값으로부터 생긴 query, key, value을 가진 multi-head attention이다. 두 번째는 position-wise feed-forward 네트워크이다. 한편, 효율적인 학습을 위해 각 하위 블록에 layer normalization과 residual connection이 도입된다. 하위 블록 입력값 x가 주어지면 출력값은 다음과 같다.

$$x + \mathrm{Subblock}\left(\mathrm{Layernorm}\left(x\right)\right). \tag{2.5}$$

디코더는 Figure 3의 오른쪽과 같다. 먼저, 학습된 character-level 임베딩을 사용하여 문자 sequence를 positional encoding이 추가된 d_{model}의 출력 인코딩으로 변환한다. 그런 다음, 그것들의 합을 $\boldsymbol{N}_d$ 디코더 블록의 stack에 입력하여 최종 디코더 출력값을 얻을 수 있다. 디코더 블록은 인코더 블록과 달리 3개의 하위 블록이 있다. 첫 번째는 동일한 query, key, value를 가지고 있는 masked multi-head attention이다. Masking은 j 번째

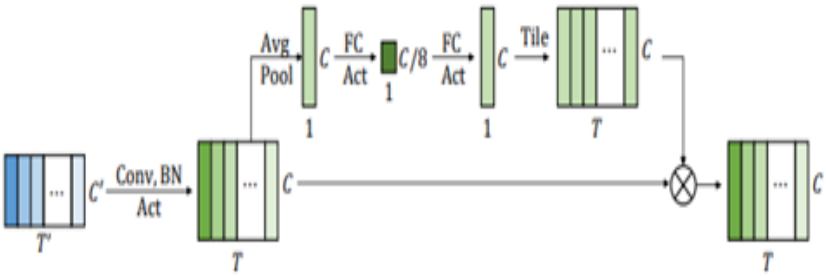


Figure 5: *1D Squeeze-and-excitation module (Han et al., 2020).*

위치에 대한 예측이 j보다 작은 위치에서의 알려진 출력값에만 의존할 수 있도록 하기 위해서이다. 두 번째는, 인코더 출력값에서 나온 key와 value, 이전 하위 블록 출력값에서 나온 query의 multi-head attention이다. 마지막은 position-wise feed-forward 네트워크이다. 인코더와 마찬가지로, layer normalization과 residual connection이 각 하위 블록에서 수행된다. 마지막으로, 디코더의 출력값은 linear projection과 subsequent softmax 함수에 의해 출력 class들의 확률값으로 변환된다.

2.4. ContextNet

자동 음성 인식에서 CNN 기반 음성인식 모델 관련 연구가 활발히 진행되고 있지만, RNN/Transformer 기반 모델보다 성능이 좋지 않았다. 성능의 차이는 global context의 부족이라고 언급하였으며, global context의 강화를 위해 squeeze-and-excitation (SE) layer를 사용한 CNN 모델 ContextNet이 등장했다 (Han 등, 2020).

Convolutional encoder가 존재하며, input sequence 음향신호 $\boldsymbol{x} = (x_1, \ldots, x_T)$를 high level representation $\boldsymbol{h} = (h_1, \ldots, h_{T'})$로 변환한다. AudioEncoder($\cdot$)는 다음과 같이 정의한다.

$$\boldsymbol{h} = \text{AudioEncoder}(\boldsymbol{x}) = C_K\left(C_{K-1}\left(\ldots C_1\left(\boldsymbol{x}\right)\right)\right). \tag{2.6}$$

$C_k(\cdot)$는 Convolution block으로, Figure 4와 같이 convolution layer가 포함되어 있으며, 각 layer에는 batch normalization, activation function이 존재한다. 또한, squeeze-and-excitation과 skip connection을 포함한다.

Squeeze-and-excitation은 local feature vector sequence를 single global context로 압축하고, 이 context를 다시 각 local feature vector로 broadcast 한 후 곱셈을 통해 둘을 병합한다. Figure 5와 같이 input x에 대해 global average pooling을 수행하여 이를 global channel-wise weight $\theta(x)$와 element-wise multiplication을 적용한 것이 SE($\cdot$) function이다. 이로 인해 naïve한 convolution layer 뒤에 SE layer를 배치할 때, convolution output에 global 정보에 대한 접근을 주게 된다.

$$\begin{aligned}
\bar{x} &= \frac{1}{T}\sum_t x_t, \\
\theta(x) &= \text{Sigmoid}\left(W_2\left(\text{Act}\left(W_1\bar{x} + b_1\right)\right) + b_2\right), \\
\text{SE}(x) &= \theta(x) \circ x.
\end{aligned} \tag{2.7}$$

$\circ$는 element-wise multiplication을 의미하며, W_1, W_2는 weight matrix를, b_1, b_2 는 bias vector를 의미한다.

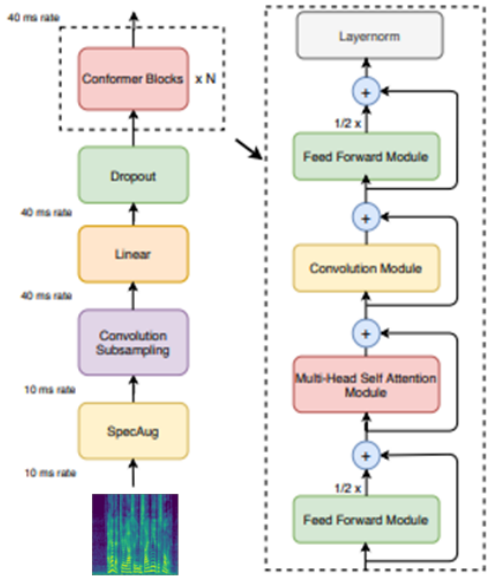


Figure 6: *Conformer encoder model architecture (Gulati et al., 2020).*

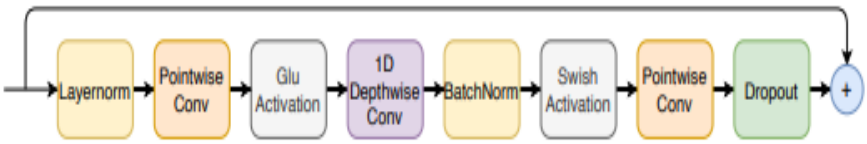


Figure 7: *Convolution module. The convolution module consists of a pointwise convolution and a 1-D Depthwise convolution with an expansion factor of 2 that projects the number of channels with the GLU activation layer. Then convolution is followed by the batch normalization and then a swish activation layer (Gulati et al., 2020).*

2.5. Conformer

Transformer 모델은 content-based global interaction을 잘 포착하지만 세분화된 local feature pattern을 추출하는 능력은 떨어진다. CNN 모델은 local feature를 효과적으로 활용하지만 많은 layer와 parameter가 필요하다는 제한이 존재한다. 2가지 모델을 동시에 활용하여 convolution-augmented transformer, 즉 conformer 모델이 등장한다 (Gulati 등, 2020).

Conformer encoder는 convolution subsampling layer를 사용해 입력을 처리하고, Figure 6과 같이 여러 conformer block을 거친다. Conformer block에는 4개의 module (feed-forward module, self-attention module, convolution module, second feed-forward module)이 존재한다.

Multi-head self-attention에 relative positional encoding을 사용하였는데, 이는 self-attention module이 다른 입력 길이에 대해 더욱 잘 일반화할 수 있으며 발화 길이의 변화에 대해 더 robust하다. 더 깊은 모델을 훈련하고, 정규화에 도움이 되는 dropout과 함께 pre-norm residual unit을 사용했다.

Convolution module은 pointwise convolution과 gated linear unit (GLU)인 gating mechanism (Wu 등, 2020)으로 시작하여, 1D depthwise convolution layer가 이어지고, batchnorm은 deep 모델 훈련을 돕기 위해 convolu-

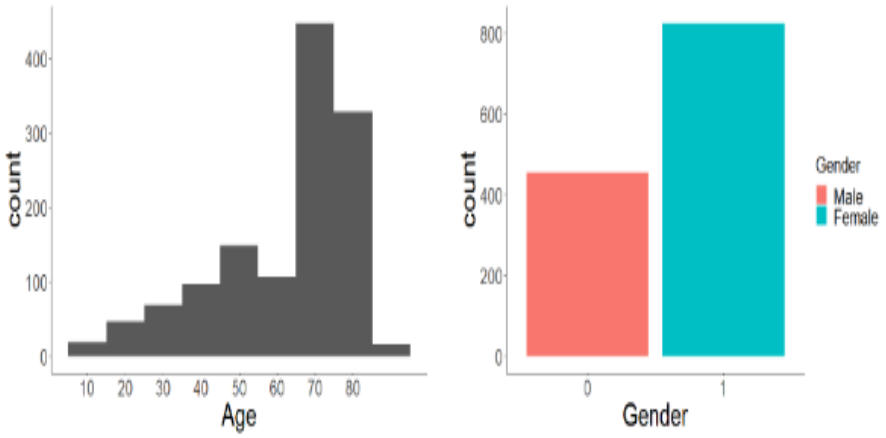


Figure 8: *Distributions of the data containing speakers' information.*

tion 직후에 위치한다. 이는 Figure 7을 통해 알 수 있다. Feed-forward module에서는 linear layer들을 통과하며, swish activation 및 dropout을 적용하여 network를 정규화하는데 도움을 준다.

Figure 7에서 conformer block은 multi-head-self-attention module과 convolution module 사이에 2개의 feed-forward module이 포함된 샌드위치 구조는 transformer block의 feed-forward layer를 2개의 half-step feed-forward layer로 대체한 Macaron-Net (Lu 등, 2019)에서 영감을 얻었다. 두 번째 feed-forward module 다음에 최종 layernorm layer가 오며 이는 다음의 식과 같다.

$$\begin{aligned}
\tilde{x}_i &= x_i + \frac{1}{2}\mathrm{FFN}\left(x_i\right), \\
x_i' &= \tilde{x}_i + \mathrm{MHSA}\left(\tilde{x}_i\right), \\
x_i'' &= x_i' + \mathrm{Conv}\left(x_i'\right), \\
y_i &= \mathrm{Layernorm}\left(x_i'' + \frac{1}{2}\mathrm{FFN}\left(x_i''\right)\right).
\end{aligned} \tag{2.8}$$

3. 실험

3.1. 데이터 설명

우리가 실험에 사용한 데이터는 언어 유창성을 기반으로 치매를 진단하기 위해 수집되었다 (Kim 등, 2014). 치매를 진단하는 일반적인 방법은 종이 기반 검사였지만, 많은 노인들이 장애가 있거나 읽기 어려운 경우 검사를 수행할 수 없기 때문에 검사에 어려움이 있다. 따라서, 이 문제를 해결하기 위해 언어 유창성에 기초한 치매 진단 방법이 제안되었다 (Chi 등, 2014). 우리는 이 데이터를 사용하여 음성 인식 실험을 수행한다.

데이터는 발화자가 동물이나 채소를 무작위로 1분동안 녹음한 음성 파일이다. 1분동안 대본을 읽지 않고 머리 속에 떠오르는 채소, 동물 등을 녹음하기 때문에 유사한 단어가 녹음된 경향이 있다. 총 데이터 크기는 3,128개이지만, 녹음 상태가 좋지 않거나 콘텐츠 부족으로 사용하기 어려운 음성 데이터를 제외한 2,939개의 데이터를 실험에 사용했다. 2,939개의 데이터 중 1,278개(43.4%)의 데이터는 나이, 성별 등 발화자의 정보가 포함되어 있다. Figure 8은 1,278개의 발화자 나이 및 성별의 분포를 보여준다. 발화자는 6세에서 88세 사이이

Table 2: Contingency table of the data

	Information	No information	Subtotal
Vegetable	9	1435	1444
Animal	1269	226	1495
Subtotal	1278	1661	2939

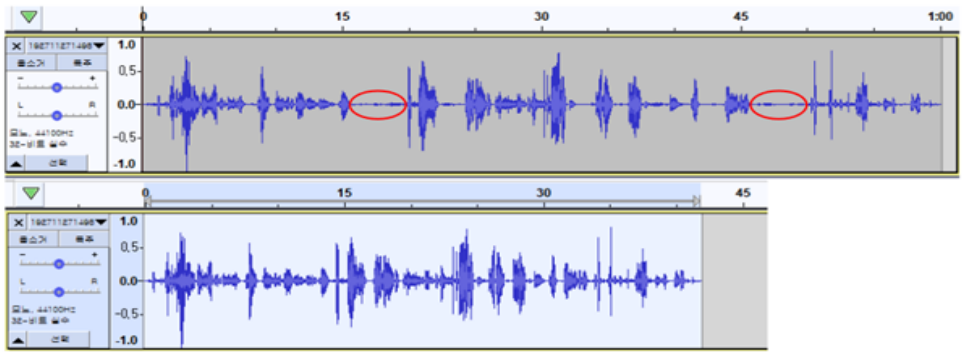


Figure 9: *Removal of silent sections.*

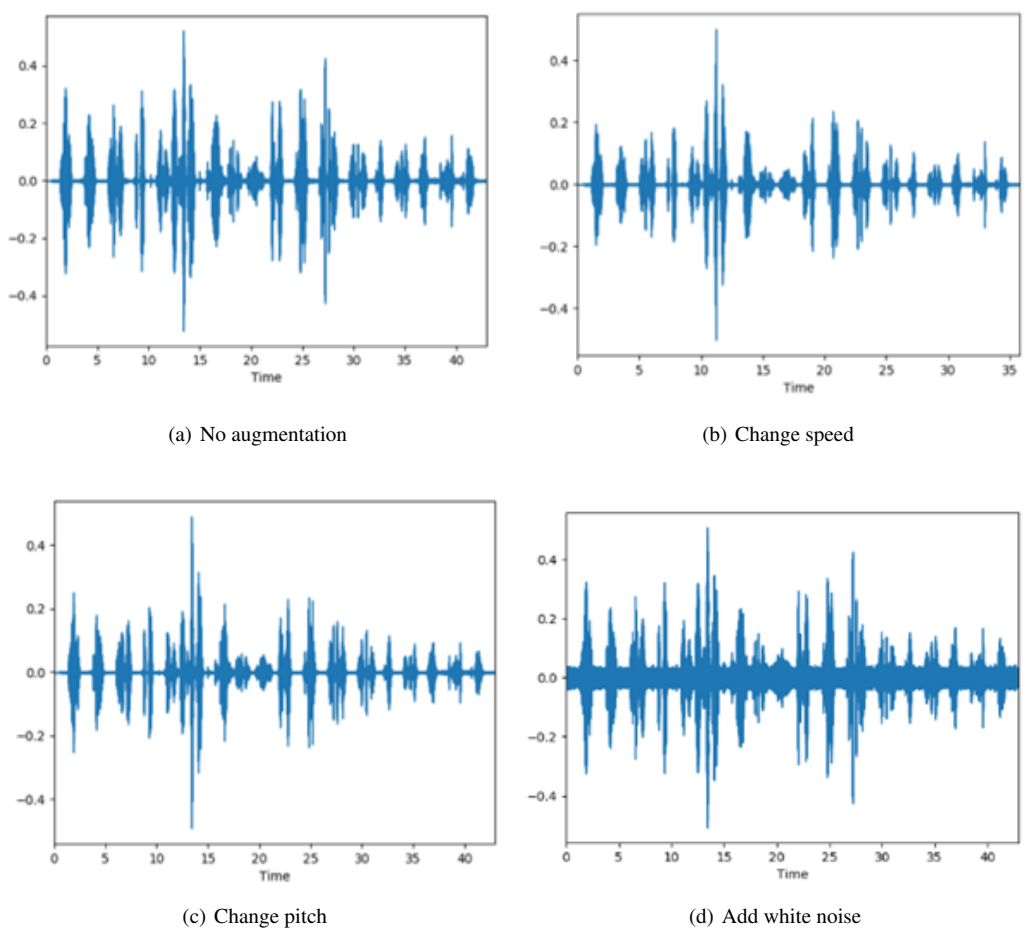
며 70대, 80대가 주를 이루고 있으며 여성이 남성보다 2배 이상 많은 것을 확인할 수 있다. 데이터는 발화자의 정보가 포함되어 있거나 없는, 동물 단어 혹은 채소 단어의 음성 파일인지 4가지의 범주로 나눌 수 있다. Table 2는 4가지 범주 관점의 데이터 분포를 보여준다. 2,939개의 데이터 중 우리는 2,642개(90%)의 데이터를 train과 validation에 사용하며, 297개(10%)의 데이터를 test에 사용한다.

녹음 중간에 침묵 구간이 존재하고 같은 단어를 반복하거나 '음', '모르겠다'와 같은 불필요한 녹음이 포함된 경우가 많다. 또한, 전반적으로 말이 느리고 발음이 정확하지 않거나 목소리가 너무 작아 인식할 수 없는 경우도 있다. 일반적인 음성 인식에서는 대화를 하거나 대본 혹은 책을 읽는, 즉 문장으로 구성된 음성 데이터를 다룬다 (Panayotov 등, 2015). 또한, 발화자는 정확한 발음으로 음성을 녹음한다. 그러나 우리의 데이터 대부분의 발화자는 노인이기 때문에 상대적으로 발음이 부정확하다. 따라서 정확한 발음으로 단어가 아닌 문장으로 구성된 open dataset이나 일반적인 음성 데이터로 학습한 pre-trained model을 우리 데이터에 적용하기 어렵다고 판단한다.

3.2. 전처리

3.2.1. Removal of silent sections and conversion of sampling rate

우리는 1분의 녹음 파일을 Audacity® recording and editing software (Audacity Team, 2020) 2.4.1 버전을 사용하여 묵음 구간을 제거했다. Figure 9를 참고하면, 각 음성의 길이는 10초에서 60초 사이로 변경되었다. 또한, 원본 데이터의 sampling rate는 44,100hz이며 우리는 16,000hz로 변경했다. Sampling rate가 너무 크고 음성 길이가 너무 길면 입력 크기도 매우 커진다. 그래서 우리의 데이터는 입력 크기가 너무 크기 때문에 sampling rate와 음성 길이를 변경했다. 예를 들어, sampling rate가 44,100hz, 음성 길이가 60초이면 input sequence는 60×44,100이지만 sampling rate가 16,000hz, 음성 길이가 10초이면 input sequence는 10×16,000이다. 우리는 이러한 방식으로 모델의 크기를 크게 줄일 수 있으며, 우리의 데이터는 매우 작기 때문에 이는 매우 중요한 요소이다.



(a) No augmentation (b) Change speed

(c) Change pitch (d) Add white noise

Figure 10: *Applied audio data augmentation.*

3.2.2. 데이터 증강 기법

묵음 구간 제거 및 sampling rate를 변경한 train & validation data 2,642개에 data augmentation을 수행한다. 일반적인 오디오 데이터의 augmentation 방법은 audio speed와 pitch를 선택한 범위 내에서 무작위로 변경하여 적용한다. 하지만 우리는 데이터에 보다 적합한 방식으로 data augmentation을 수행한다.

첫째, 노인들의 말속도는 일반적으로 느리기 때문에 (Lee와 Gwon, 2014), Figure 10의 (b)와 같이 원래보다 1.2배 빠르게 속도를 증가했다. 둘째, 우리 데이터의 대부분은 pitch가 낮기 때문에, 실제 음성을 들을 때 우리는 음성을 잘 이해할 수 없다. 그래서 Figure 10의 (c)와 같이 pitch를 약간 증가시켰다. Pitch를 증가시키는 것은 pitch를 무작위로 변경하는 것보다 모델 학습에 있어서 더 좋은 성능을 보인다. 셋째, Figure 10의 (d)와 같이 white noise를 추가했다.

이러한 3가지 data augmentation 방법을 통해 데이터셋의 개수를 4배(10,568개)만큼 증가시켰다. 10,568개의 데이터 중 8,454개(80%)는 training set, 2,114개(20%)는 validation set으로 사용했다.

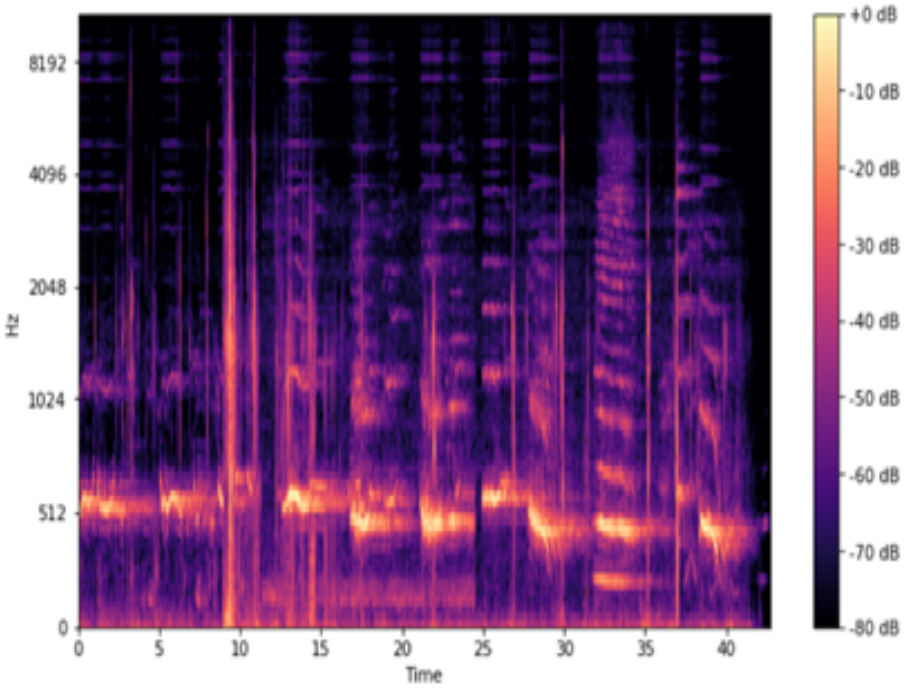


Figure 11: *Mel-spectrogram plot.*

3.2.3. Mel-spectrogram

Mel-spectrogram이란 시간에 따라 주파수 특성이 달라지는 오디오를 분석하기 위한 특징 추출 기법 중 하나이다. 사람들의 귀는 음성 신호를 인식할 때 주파수를 linear scale로 인식하지 않으며, 낮은 주파수에서의 변화를 높은 주파수에서의 변화보다 더 예민하게 받아들인다. 그래서 인간이 이해하기 힘든 오디오 spectrogram의 y 축 주파수를 mel-scale로 변환한 것이 mel-spectrogram이다. 이는 Figure11과같다. Mel-scale이란 pitch에서 발견한 사람의 음을 인지하는 기준을 반영한 scale 변환 함수이다. 변환하는 식은 아래와 같으며, f는 주파수(frequency)를 의미한다.

$$\mathrm{Mel}(f) = 2595 \log\left(1 + \frac{f}{700}\right). \tag{3.1}$$

3.3. 결과

우리는 우리의 데이터를 사용하여 여러 End-to-end ASR 모델을 학습했다. CER을 통해 모델의 성능을 비교, 측정하였으며 google web speech API 및 clova speech recognition (CSR)을 통해 얻은 텍스트 결과 또한 비교했다. CER은 insertions (I), substitutions (S), deletions (D)의 합을 출력 문자의 총 길이(n)로 나눈 값이다.

$$\mathrm{CER} = \frac{(I + S + D)}{n}. \tag{3.2}$$

Table 3: Performance comparison of models

Model	CER(%)
Google API	61.16
Naver API	63.21
Listen Attend and Spell (LAS)	36.24
Speech-transformer	22.21
Contextnet	13.17
Conformer	15.69

Table 4: Comparison of the data augmentation effect in ContextNet model

Kind of augmentation	CER(%)
No augmentation (NA)	34.74
Add white noise (WN)	26.33
Change pitch (CP)	27.02
Change speed (CS)	23.90
WN + CP	23.58
WN + CS	21.35
CP + CS	23.46
WN + CP + CS	13.17

CER 값이 작을수록 모델이 더 우수하다. 한국어의 경우, 영어와 달리 여러 모음, 자음의 조합으로 무수히 많은 단어쌍들이 존재하기 때문에 WER보다는 CER을 성능 평가의 척도로 사용하는 것이 일반적이다.

Google ASR 모델의 경우 google API를 통해 음성 데이터를 입력으로 사용하여 출력 문자를 생성한다. 그 후, CER은 각 음성 데이터의 출력 문자를 사용하여 더한 후 평균화하여 계산되었다. Naver ASR 모델의 경우, google과 비슷한 방식으로 naver papago API를 통해 각 데이터의 문자를 출력하여 CER을 획득했다.

Table 3은 실험의 결과를 보여준다. API는 일반적인 훈련 데이터, 즉 발음이 좋은 성인들의 음성으로 학습되어 있기 때문에 60세 이상의 발음이 좋지 않은 노인들로 구성된 우리 데이터에는 좋은 성능을 보여주지 못한다. 반면에, 최근 state-of-the-art (SOTA) 성능을 보였던 ContextNet과 Conformer 모델은 각각 CER(%)값이 13.17%, 15.69%의 성능을 보인다. ContextNet 모델이 가장 좋은 성능을 보여주기 때문에 우리의 데이터에 가장 적합한 모델은 ContextNet 모델임을 알 수 있다.

3.4. Ablation study

좋은 성능을 보여준 두 모델, ContextNet과 Conformer 모델에 한하여 우리가 적용한 data augmentation의 효과를 입증하기 위해 실험을 진행했다. 실험에서 사용한 hyper parameter는 모든 data augmentation을 적용했을 때의 hyper parameter를 동일하게 적용했다. 총 7가지 경우의 수를 실험하였으며, NA는 data augmentation을 적용하지 않았을 때, WN은 white noise만 적용했을 때, CP는 change pitch만 적용했을 때, CS는 change speed만 적용했을 때, WN + CP는 white noise와 change pitch를 적용했을 때, WN + CS는 white noise와 change speed를 적용했을 때, CP + CS는 change pitch와 change speed를 적용했을 때를 의미한다.

Table 4는 ContextNet 모델의 data augmentation 효과를 비교하였으며, change speed만을 적용한 모델이 no augmentation보다 CER(%)값이 약 10% 이상의 성능 향상을 보여준다. 또한, WN + CS 모델이 NA 모델보다 13% 이상의 성능 향상을 보여주며, CS만 적용한 모델보다 좋은 성능을 보이는 것을 확인할 수 있다. 모든 data augmentation을 적용했을 때, ContextNet 모델은 CER(%)값이 13.17%의 성능을 보이며 이는 아무런 data

Table 5: Comparison of the data augmentation effect in Conformer model

Kind of augmentation	CER(%)
No augmentation (NA)	29.76
Add white noise (WN)	24.58
Change pitch (CP)	25.00
Change speed (CS)	23.76
WN + CP	21.35
WN + CS	22.42
CP + CS	21.39
WN + CP + CS	15.69

augmentation을 적용하지 않았을 때(NA)보다 20% 이상의 성능 향상을 보여준다. Table 5는 Conformer 모델의 data augmentation 효과를 비교하였으며, change speed만을 적용한 모델이 no augmentation보다 CER(%)값이 약 6% 이상의 성능 향상을 보여준다. 또한, WN + CP 모델이 NA 모델보다 8% 이상의 성능 향상을 보여주며, WN + CP + CS 모델은 CER(%)값이 15.69%의 성능을 보이며 이는 NA 모델보다 14% 이상의 성능 향상을 보여준다. 이를 통해 우리는 data augmentation이 모델의 성능 향상에 중요한 요인임을 보여준다. 또한, 우리의 데이터 대부분의 발화자는 노인으로, 말이 느리다는 특성 때문에 change speed 방법이 다른 data augmentation 방법보다 성능 향상에 더 효과적인 것을 확인할 수 있다.

4. 결론

이 연구는 우리의 데이터를 사용하여 여러 음성인식 모델을 비교했다. 우리의 데이터는 60세 이상의 노인들이 많으며, API가 일반적인 성인 음성에 초점이 맞춰져 있기 때문에 우리의 실험은 노인의 목소리가 젊은 사람들의 목소리와 다르다는 것을 보여준다. 노인의 목소리가 일반 성인의 목소리와 비슷하다면 우리가 직접 실험한 모델은 API와 비교하여 성능에 큰 차이가 없을 것이다. 그동안 음성 인식에 사용되는 흔한 데이터셋을 사용한 대부분의 실험은 발화자의 나이 또는 기타 정보 등을 고려하지 않았다. 그러나 우리의 연구에 따르면 노인의 목소리는 일반 성인의 목소리와 다르지만, 그 차이점이 정확히 무엇인지 알 수 없다. 앞으로의 음성 인식 연구에서는 발화자의 정보 또는 특성 등을 고려한다면 더 좋은 결과를 얻을 수 있다고 생각한다.

우리는 일반적인 자동 음성 인식에서 다루는 대화 또는 대본 문장을 읽는 데이터가 아닌 단어에 대한 자동 음성 인식 문제를 연구했다. 대부분의 발화자는 치매환자를 포함한 60세 이상의 노인이였으며, 따라서 일반적인 성인들에 비해 발음이 정확하지 않다. 기존 자동 음성 인식 문제와 다른 형태의 데이터이기 때문에 우리 데이터를 통해 기존 음성 인식을 새로 학습해야 했다. SpecAugment, 묵음 구간 제거, 일반적인 오디오 데이터 증강을 적용하여 google과 naver의 API에 비해 약 45% 이상 향상된 ContextNet 모델과 Conformer 모델이 우리의 데이터에 적합하다는 것을 알았다.

치매 환자를 포함한 노인 한국어 음성 데이터에 대한 실험을 했지만, 모든 발화자가 치매 환자는 아니다. 그래서 우리는 치매 환자의 목소리 특성을 고려하지 않았다. 우리 데이터의 음성 인식률은 API에 비해 괜찮지만, 일반적인 음성 데이터의 실험보다 인식률이 낮다. 앞으로의 연구에서는 치매 환자, 노인 및 일반 성인 그룹을 그룹별로 비교하고 치매 환자의 차이를 비교하는 실험을 할 것이다. 또한, 치매 환자 음성의 특성에 대한 이전의 연구들을 참고하여, 치매 환자의 음성 특징을 고려한 자동 음성 인식을 연구할 것이다. 마지막으로, 치매 진단이 어떻게 수행되는지를 파악하고, 딥러닝 기반 자동 음성 인식 모델을 바탕으로 치매 진단 가능성을 파악하고, 치매에 대한 진단을 실시할 예정이다.

References

Malik M, Malik MK, Mehmood K, and Makhdoom I (2020). Automatic speech recognition: A survey, *Multimedia Tools and Applications*, **80**, 9411–9457.

Kim YG and Lee JW (2020). Development of Korean automatic speech recognition model using transformer, *Proceedings of Symposium of the Korean Institute of Communications and Information Sciences*, **71**, 659–660.

Lee SG and Kwon SI (2014). Elderly speech analysis for improving elderly speech recognition, *Communications of the Korean Institute of Information Scientists and Engineers*, **32**, 16–20.

Young V and Mihailidis A (2010). Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review, *Assistive Technology*, **22**, 99–112.

Audacity Team, Audacity® (2020). Free audio editor and recorder [Computer application] Version 2.4.1. 2020.

Chi YK, Han JW, Jeong HJ *et al.* (2014). Development of a screening algorithm for Alzheimer's disease using categorical verbal fluency, *PLOS ONE*, **9**, e84111.

Chakraborty K, Talele A, and Upadhya S (2014). Voice recognition using MFCC algorithm, *International Journal of Innovative Research in Advanced Engineering (IJIRAE)*, **1**, 2349–2163.

Chan W, Jaitly N, Quoc VL, and Vinyals O (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition, In *proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, 4960–4964.

Dong L, Xu S, and Xu B (2018). Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition, In *proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, 5884–5888.

Han W, Zhang Z, Zhang Y *et al.* (2020). Contextnet: Improving convolutional neural networks for automatic speech recognition with global context, Available from: arXiv preprint arXiv:2005.03191..

Wu Z, Liu Z, Lin J, Lin Y, and Han S (2020). Lite transformer with long-short range attention, Available from: arXiv preprint arXiv:2004.11886, 2020.

Lu Y, Li Z, He D, Sun Z, Dong B, Qin T, Wang L, and Liu TY (2019). Understanding and improving transformer from a multi-particle dynamic system point of view, Available from: arXiv preprint arXiv:1906.02762.

Gulati A, Qin J, Chiu CC *et al.* (2020). Conformer: Convolution-augmented transformer for speech recognition, Available from: arXiv preprint arXiv:2005.08100.

Graves A and Jaitly N (2013). Towards end-to-end speech recognition with recurrent neural networks, *Proceedings of the 31st International Conference on International Conference on Machine Learning PMLR*, **32**, 1764–1772.

Kim KW, Ji YG, and Han JW (2014). inventors. Method of diagnosing dementia based on verbal fluency and apparatus therefore. South Korea patent KR101437569B1, 2014 Sep 04.

Panayotov V, Chen G, Povey D, and Khudanpur S (2015). Librispeech: An ASR corpus based on public domain audio books, In *proceedings of 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, 5206–5210.

Park DS, Chan W, Zhang Y, Chiu C, Zoph B, Cubuk ED, and Quoc VL (2019). Specaugment: A simple data augmentation method for automatic speech recognition, Available from: arXiv preprint arXiv:1904.08779

Received October 10, 2022; Revised December 11, 2022; Accepted December 13, 2022

치매 환자를 포함한 한국 노인 음성 데이터 딥러닝 기반 음성인식

문정현[*a], 강준서[*a], 김기웅[bcd], 배종빈 [bc], 이현준[e], 임창원[1,a]

[a]중앙대학교 응용통계학과; [b]서울대학교병원 정신건강의학과; [c]서울대학교 정신의학과;
[d]서울대학교 뇌인지과학과; [e]세븐포인트원

요 약

본 연구에서는 발화자가 동물이나 채소와 같은 일련의 단어를 무작위로 일 분 동안 말하는 한국어 음성 데이터에 대한 자동 음성 인식(ASR) 문제를 고려하였다. 발화자의 대부분은 60세 이상의 노인이며 치매 환자를 포함하고 있다. 우리의 목표는 이러한 데이터에 대한 딥러닝 기반 자동 음성 인식 모델을 비교하고 성능이 좋은 모델을 찾는 것이다. 자동 음성 인식은 컴퓨터가 사람이 말하는 말을 자동으로 인식하여 음성을 텍스트로 변환할 수 있는 기술이다. 최근 들어 자동 음성 인식 분야에서 성능이 좋은 딥러닝 모델들이 많이 개발되어 왔다. 이러한 딥러닝 모델을 학습시키기 위한 데이터는 대부분 대화나 문장 형식으로 이루어져 있다. 게다가, 발화자들 대부분은 어휘를 정확하게 발음할 수 있어야 한다. 반면에, 우리 데이터의 발화자 대부분은 60세 이상의 노인으로 발음이 부정확한 경우가 많다. 또한, 우리 데이터는 발화자가 1분 동안 문장이 아닌 일련의 단어를 무작위로 말하는 한국어 음성 데이터이다. 따라서 이러한 일반적인 훈련 데이터를 기반으로 한 사전 훈련 모델은 본 논문에서 고려하는 우리 데이터에 적합하지 않을 수 있으므로, 우리는 우리의 데이터를 사용하여 딥러닝 기반 자동 음성 인식 모델을 처음부터 훈련한다. 또한 데이터 크기가 작기 때문에 일부 데이터 증강 방법도 적용한다.

주요용어: 한국 노인 음성 데이터, 자동 음성 인식, 딥러닝, 데이터 증강

이 논문은 2021년도 중앙대학교 CAU GRS 지원에 의하여 작성되었음.
이 성과는 과학기술정보통신부의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2021R1F1A1056516).
[*]**이 저자들은 공동 제 1저자들이다.**
[1]교신저자: (06974) 서울특별시 동작구 흑석로 84, 중앙대학교 응용통계학과. E-mail: clim@cau.ac.kr