JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 24 No. 12, pp. 17-24, December 2019
https://doi.org/10.9708/jksci.2019.24.12.017

An Automatic Data Construction Approach for Korean Speech Command Recognition

Yeonsoo Lim*, Deokjin Seo*, Jeong-sik Park**, Yuchul Jung*

*Student, Dept. of Computer Engineering, Kumoh National Institute of Technology, Gumi, Korea
*Student, Dept. of Computer Engineering, Kumoh National Institute of Technology, Gumi, Korea
**Professor, Dept. of English Linguistics & Language Technology, Hankuk University of Foreign Studies, Seoul, Korea
*Professor, Dept. of Computer Engineering, Kumoh National Institute of Technology, Gumi, Korea

[Abstract]

The biggest problem in the AI field, which has become a hot topic in recent years, is how to deal with the lack of training data. Since manual data construction takes a lot of time and efforts, it is non-trivial for an individual to easily build the necessary data. On the other hand, automatic data construction needs to handle data quality issue. In this paper, we introduce a method to automatically extract the data required to develop Korean speech command recognizer from the web and to automatically select the data that can be used for training data. In particular, we propose a modified ResNet model that shows modest performance for the automatically constructed Korean speech command data. We conducted an experiment to show the applicability of the command set of the health and daily life domain. In a series of experiments using only automatically constructed data, the accuracy of the health domain was 89.5% in ResNet15 and 82% in ResNet8 in the daily lives domain, respectively.

▸ **Key words**: Korean Speech Command, Speech Recognition, Automatic Data Construction, ResNet, CNN

[요 약]

최근 화두가 되고 있는 AI분야에서 가장 큰 문제점은 학습데이터의 부족 문제를 꼽을 수 있다. 수동 데이터 구축에는 많은 시간과 노력이 소요되기에 개인이 손쉽게 필요 데이터를 구축하기는 매우 어렵다. 반면, 수동 데이터 구축에 비해 자동으로 구축하는 것은 높은 품질을 유지하는 것이 관건이다. 본 논문에서는 한국어 음성 명령어 인식기 개발에 필요한 데이터를 웹에서 자동으로 추출하고, 학습데이터로 사용할 수 있는 데이터를 자동으로 선별하는 방법을 소개한다. 특히, 자동 구축된 한국어 음성 데이터를 대상으로 우수한 성능을 보이는 ResNet기반의 수정 모델을 기반으로, 건강 및 일상생활도메인의 명령어 셋을 대상으로 적용가능성을 보이기 위한 실험을 진행하였다. 자동으로 구축된 데이터만을 사용한 일련의 실험에서 건강도메인은 ResNet15에서 89.5%, 일상생활도메인에서는 ResNet8에서 82%의 정확도를 보임으로써, 자동 수집 데이터의 활용 가능성을 검증하였다.

▸ **주제어**: 한국어 명령어 인식, 음성인식, 자동 데이터 구축, 레스넷, 합성곱신경망

• First Author: YeonSoo Lim, Corresponding Author: Yuchul Jung
*Yeonsoo Lim(yslim6168@kumoh.ac.kr), Dept. of Computer Engineering, Kumoh National Institute of Technology
*Deokjin Seo(406023@naver.com), Dept. of Computer Engineering, Kumoh National Institute of Technology
**Jeong-sik Park(parkjs@hufs.ac.kr), Dept. of English Linguistics & Language Technology, Hankuk University of Foreign Studies
*Yuchul Jung(jyc@kumoh.ac.kr), Dept. of Computer Engineering, Kumoh National Institute of Technology
• Received: 2019. 11. 26, Revised: 2019. 12. 18, Accepted: 2019. 12. 20.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

최근 IT분야에서 이슈가 되고 있는 기술 중 하나는 음성인식 기술이다. 음성인식은 사람의 발화가 담긴 음성 데이터를 학습하여 컴퓨터가 해석해내는 것인데, 최근 딥러닝을 이용한 음성인식 고도화 연구가 활발하다. 하지만 여전히 음성인식에 필요한 충분한 데이터를 확보함에 있어 제약이 많으며, 텍스트분야 혹은 영상분야의 데이터를 구축하는 것과 비교하여 그 구축과정이 더 까다롭다고 알려져 있다. 최근에는 YouTube나 Ted등 온라인에 적재된 동영상을 통하여 음성인식용 데이터를 자동으로 구축하는 연구가 활발하다. 수동 구축에 비해 온라인상의 콘텐츠들로부터 자동으로 데이터를 수집하는 것은 구축비용 절감이라는 큰 장점이 있다. 최근 온라인상의 동영상들을 이용하여 음성 데이터 셋을 자동으로 구축하고 음성인식의 성능향상을 시도하는 연구가 등장하고 있는데, 대표적으로 Lakomkin et al.[1]에서는 YouTube를 통해 자동으로 수집한 데이터를 활용하여 복수개의 음성데이터 셋에 대한 성능향상을 확인하였다.

대다수의 음성인식 연구가 영어권에서 수행되었으며, 특히 음성 명령어 (Speech Command) 같은 경우, 영어권 데이터 구축에 대한 연구가 몇몇 있긴 하지만, 한국어 음성 명령어 데이터 구축에 대한 연구는 거의 없다. 또한, 기존의 연구들 중 한국어 음성 명령어 데이터를 자동 구축하고 이를 딥러닝 기법으로 활용 가능한지에 대해 분석한 사례는 극히 드물다.

음성 명령어의 인식에는 다양한 기계학습 기반의 알고리즘이 사용될 수 있는데, 기존 HMM기반의 KeyWord Spotting (KWS) 알고리즘 및 최근 딥러닝 기반의 CNN, DS-CNN, Res8, Res15, TC-ResNet 등의 알고리즘들이 소개된바 있다.

본 연구에서는 YouTube 동영상에서 음성 데이터 추출을 통해 건강 도메인과 일상생활 도메인의 음성 명령어 인식에 쓰일 수 있는 한국어 데이터를 자동으로 수집·정제하는 방법을 제시하고, 총 30개의 명령어에 대해 1,102 건의 학습데이터 (건강: 15개/550건, 일상생활: 15개/552건)를 구축하였다. 또한, 음성명령어 인식에 효과적인 딥러닝 알고리즘인 ResNet8, ResNet15, CNN 등을 활용한 실험을 진행하여, 자동으로 구축된 데이터만을 사용한 일련의 실험에서 건강도메인은 ResNet15에서 89.5%, 일상생활도메인에서는 ResNet8에서 82%의 정확도를 보임으로써, 자동수집 데이터의 활용 가능성을 검증하였다.

II. Preliminaries

1. Related works

1.1 Data Construction for Speech Recognition

높은 성능의 음성인식을 위해서는 많은 양의 음성데이터가 필요하고, 이는 많은 비용과 인적 리소스를 필요로 한다. 그 구축방법에서는 크게 수동 구축과 자동구축으로 나눠볼 수 있다.

1.1.1 Manual Data Construction

현재까지 대부분의 음성인식 실험은 데이터를 수동으로 구축하였다. 직접 발화자, 문장, 단어 등에 대해 음성 녹음을 이용하여 데이터를 구축할 수 있는데, 이들 중 공개된 데이터들은 제공기관 또는 웹 사이트를 통해 유/무료로 데이터를 획득할 수 있다. 무료로 한국어 음성 데이터 셋을 제공하는 대표적인 사이트로는 Zeroth [2], KSS [3] 등이 있다. Zeroth의 한국어 데이터 셋을 보면 76.6시간 분량, 35139개의 표현, 137명의 발화자, 16472개의 문장으로 이루어진 데이터를 제공하고 있으며, KSS에서 제공하는 데이터 셋을 단일 여성 화자의 12시간 분량, 12853개의 표현을 제공하고 있다. 수동으로 데이터 셋을 직접 구축 방법 외에 기 구축된 음성데이터를 유/무료로 획득하는 방법이 있지만, 비용이 발생할 수도 있고, 경우에 따라 자신이 원하는 데이터 셋이 아닌 경우가 있을 수 있다.

1.1.2 Automatic Data Construction

Lakomkin et al.[1]에서는 YouTube의 방대한 음성과 자막을 이용하여 많은 양의 고품질의 음성데이터를 자동으로 구축하여 인식률 개선을 보였다. 반면 Kaewprateep et al.[4]에서는 적은 양의 데이터를 구축하여 음성인식의 성능을 확인한 사례도 있었지만, 높은 성능을 보이지는 않았다.

Choi et al.[5]에서는 고성능의 ASR(Automatic Speech Recognition)을 위해 TED 혹은 TEDx와 같은 웹에 존재하는 음성데이터로부터 음성 데이터베이스를 반자동으로 구축하였다. 이 경우 한국어 음성데이터와 한국어 자막 데이터를 자동 생성하였지만, 녹음의 부정확성, 발음의 모호성, 비 이상적인 오디오 조건으로 인한 저품질 레코딩 등을 극복하기 위해, 자원봉사자들을 동원하여 데이터 검증을 추가적으로 진행하였다.

1.2 Data Augmentation of Speech Recognition

이미지 인식 등 타 패턴인식 분야에 이어 음성인식에서도 변형된 음성에 강인한 모델을 만들기 위한 증강(augmentation) 기법을 적용하게 되었다.

최근 Zhang et al.[6]에서는 Mixup기반 증강기법을 소개하였는데 이는 두 데이터의 가공 전(raw) 입력과 원-핫(one-hot) 임베딩된 라벨을 가중치 선형 보간법(weighted-linear- interpolation)을 활용하여 새로운 샘플(sample)을 생성하는 기법이다. 음성 명령어 데이터 셋을 이용한 실험에서 적절한 가중치를 적용한 경우, LeNet은 ERM (Empirical Risk Minimization)을 사용했을 때보다 WER (Word Error Rate) 이 9.8%에서 10.1%로 증가하였고, 모델의 capacity가 더 큰 VGG-11에서는 WER이 5.0%에서 4.0%로 감소하는 것을 확인하였고, 증강기법의 효용성을 보였다.

1.3 Algorithms of Automatic Speech Recognition

KWS (KeyWord Spotting) [7-8] 알고리즘은 음성인식에서 입력받은 전체 문장 중에서 중요한 키워드 또는 필요한 키워드만을 가려내는 음성인식 방법이다. 데이터를 정제/가공함에 있어서 사람이 수작업을 하지 않는 (hands-free) 인터페이스를 제공하는 잠재적인 기술로 유명한 알고리즘이다. 과거에는 GMM으로 음소를 모델링하고 이 음소들의 연속적 변화를 HMM으로 예측한 연구들이 많이 진행되었지만, 불과 몇 년 사이에 딥러닝으로 대체되었다.

특히 Chen et al.[9] 에서는 일상 대화 말뭉치들을 3,000시간 녹음한 음성데이터와 특정 명령어를 발화한 음성데이터로 부터 학습한 HMM 모델보다 딥 뉴럴 네트워크 기반의 KWS 성능이 False Reject Rate측면에서 45% 개선되는 것을 보여준다.

Warden [10]에서는 KWS모델을 학습하고 평가할 음성 명령어 데이터 셋 구축을 위해, 18811명의 화자들을 대상으로 64,727개의 명령어 발화를 직접 녹음하였다. 또한 노이즈 합성과 발화에서 소리가 큰 부분을 추출하는 등 데이터의 품질을 올리기 위한 방법을 제안하였고, CNN기반의 단순한 딥러닝기법을 이용하여 85.4%의 성능을 보였다.

최근 ResNet구조 기반의 연구 Tang et al.[11]에는 ResNet15모델을 따르는 Deep Residual Learning을 사용하여 Warden [10]의 데이터 셋에서 95.8%의 정확도를 증명함으로써, ResNet 구조의 모델이 음성명령어 인식에서 우수한 성능을 보이는 것을 검증했다.

따라서, 본 논문의 실험들에서는 음성명령어 인식에서 성능이 입증된 ResNet구조의 모델을 사용하여 구축된 한국어 음성 명령어 인식 성능 테스트를 진행하였다.

III. The Proposed Scheme

1. Data Construction

본 실험의 데이터는 실험 도메인과 질의어, 질의어의 수량을 선정 한 뒤, 선정된 질의어와 일치하는 음성 발화 부분을 가지는 동영상을 Google에서 제공하는 YouTube Search API (**https://cloud.google.com/speech-to-text/**)를 이용하여 검색한다. 검색된 YouTube 영상에서 자막 부분과, 질의어에 해당 하는 음성 부분을 수집한다. 동영상 내에 음성 발화 부분은 1초 이내의 길이를 가지며, 자막 (Script text)을 형태소 단위로 분석하여 각 명사에 대응 되는지를 확인한 후 일치하는 음성부분을 추출한다.

위의 과정을 바탕으로 진행한 실험에서 약 3일간의 데이터 수집을 통해 구축된 데이터는 Table 1과 같다. 명령어 도메인은 건강과 (스마트폰을 사용하면서 나올만한) 일상생활 등의 두 부류로 나누었다. 음성데이터 자동 수집을 통해, 건강 도메인에서는 명령어 세트가 4026개, 명령어에 해당하는 파일들의 합이 20172개의 데이터가 구축되었고, 일상생활 도메인에서는 명령어 세트가 4558개, 명령어에 해당하는 파일들의 합이 21853개가 구축되었다.

2. Data Refinement

본 연구에서는 수집된 데이터 중 실험에 사용 가능한 데이터를 선별하기 위해 Google Cloud에서 제공하는 STT (Speech-to-Text) API를 이용하여 정제 하였다. YouTube로부터 구축한 데이터 중 상당수가 도메인의 명령어와 맞지 않는 단어를 발화하고 있기 때문에 명령어와 STT의 출력과 비교하여 올바른 발화를 나타내는 데이터만 사용하였다. 이 중 건강 도메인에서는 명령어 세트가 354개, 명령어에 해당하는 파일들의 합이 2044개의 데이터가 정제되었고, 일상생활 도메인에서는 명령어 세트가 374개, 명령어에 해당하는 파일들의 합이 2039개의 데이터가 정제되었다 (Table 1).

Table 1. Generated data & Preprocessing data

Domain	YouTube API		STT Filtering		Final Data	
	CMD	Files	CMD	Files	CMD	Files
Healthcare	4026	20172	354	2044	15	550
Daily lives	4558	21853	374	2039	15	552
CMD: Numbers of Speech Commands Files: Number of Raw Files						

이어서, 각 도메인에 적합한 명령어 세트들로 구성하여 15개씩만 남도록 한 번 더 정제하였으며, 결과적으로 건강 도메인에서 명령어 15개와 음성 파일 550건, 일상생활 도

메인에서 명령어 15개와 음성 파일 552건을 획득하였다.

Table 2. Details of final experiment data

Healthcare		Daily lives	
Speech Commands	Files	Speech Commands	Files
vitamin	128	naver	131
virus	80	calender	68
painkiller	61	e-mail	53
diet	38	camera	44
suddenly	37	youtube	37
side-effect	28	service	31
protein	25	image	26
magnesium	23	story	23
aspirin	23	fine dust	23
energy	20	navigation	20
tylenol	20	coffee shop	20
caffeine	19	mic	20
stress	17	kakao	20
hypertension	16	message	18
endoscope	15	internet	18

3. Feature Extraction

특징 추출 알고리즘으로는 MFCC (Mel-Frequency Cepstrum Coefficient)를 이용하였다. MFCC는 인간의 주파수 특성과 청각시스템을 고려한 음성인식에서 대표적으로 사용하는 벡터이며, 일반적으로 13차원의 계수를 추출한다.

본 실험에서는 원래의 음성 데이터 입력을 1024~2048개의 Fourier transform, 512의 hop, 16kHz의 sampling rate 기준으로 40차원의 MFCC 계수를 추출하였다.

4. Data Augmentation

데이터에 Mixup기법을 적용하여 데이터 특징을 증강시켰다. 두 음성 데이터와 원-핫 임베딩된 라벨을 가중치 선형 보간법(weighted-linear-interpolation)을 활용하여 새로운 데이터를 구성할 수 있고 이는 특징 증강효과를 의미한다.

본 실험에서는 디리클레 분포의 Alpha = 2, Beta = 2로 나온 random한 값을 mixup의 parameter인 alpha의 값으로 실험 하였다. alpha로 특징과 라벨들 사이 선형 보간법의 강도를 정한다.

5. Model Architecture

5.1 ResNet

본 실험에서는 이미지 분류에서 높은 성능을 보여준 ResNet 기반의 모델 [12-13]을 사용하였다. Residual Network의 기본 구조는 블록의 입력을 함수 F(x)와 입력의 Identity를 함수 H(x)의 합에 대하여 활성화 함수인 ReLU를 한 결과이다.

예를 들어, Fig. 1과 Fig. 2에 볼 수 있듯이, 함수 F(x)의 결과와 함수 H(x)의 결과를 합으로 표현되나 실제로는 활성화 함수 f(x)로 인해 f(h(x) + F(x))가 된다.

본 실험에서 사용한 ResNet15와 ResNet8은 2개의 컨볼루션 레이어를 쌓은 함수 F(x)를 사용하였는데, 이 때 입력이 두개의 레이어를 거치게 되면 미분 값이 너무 작아 문제를 일으킬 수 있다 [11].

이를 방지하기 위해 1개의 컨볼루션 레이어를 사용한 H(x)의 값을 더해 미분 값을 적어도 1이상의 값이 나올 수 있도록 하여 학습이 제대로 일어나지 않는 현상을 최소화하였다.

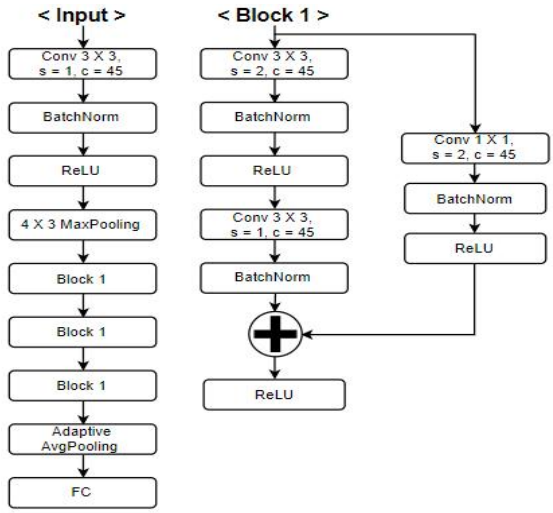

Fig. 1. ResNet8

원래 ResNet8 [14]에서는 함수 H(x)를 입력 그 자체의 Identity를 사용하게 되는데 본 실험에서는 성능을 조금 더 개선하고자 ResNet15와 같이 1개의 컨볼루션 레이어를 사용한 함수 H(x)를 사용하였다.

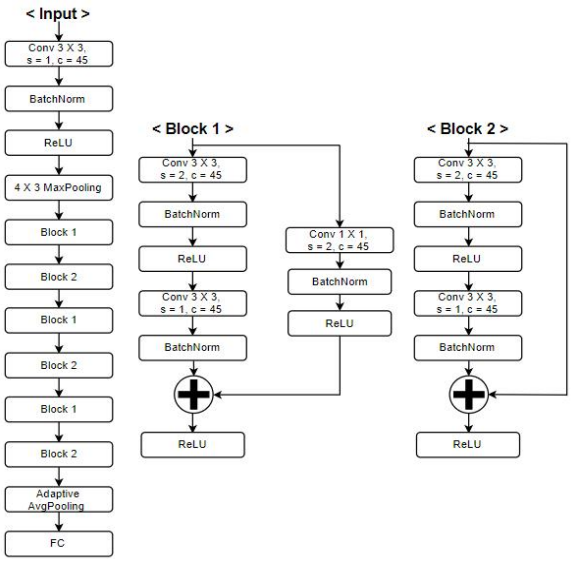

Fig. 2. ResNet15

5.2 Basic Convolutional Neural Net (CNN)

최근 성능이 좋은 Convolution Neural Network (CNN)의 구조는 점점 깊은 구조를 채택하고 있다. Tang et al.[11]에서는 2개의 컨볼루션 레이어와 1개의 Deep Neural Network (DNN)를 사용하는 cnn-trad-fpool3 모델을 KWS에 최초로 사용하였다. 이는 대다수의 KWS의 실험에서 베이스라인으로 활용된다.

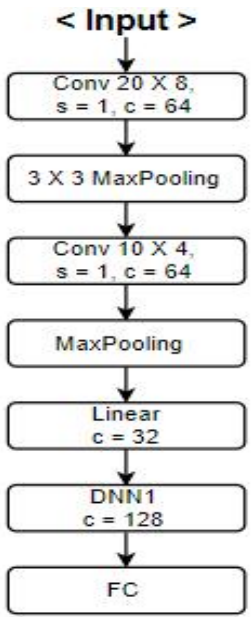

Fig. 3. Cnn-trad-fpool3

6. Experiments

학습을 위해 수집된 데이터의 80%를 학습 셋(training set), 10%를 검증 셋(validation set), 그리고 나머지 10%를 테스트 셋(testing set)으로 사용하였다. 건강 도메인과 일상생활 도메인 각각 440개와 442개를 학습 셋으로 사용했고, 검증 셋과 테스트 셋은 두 도메인 모두 55개로 동일하게 구성하였다.

특정 단어를 발화하는 부분만을 기계적으로 선별 추출하여 MFCC 특징을 추출한다. Fig. 4는 “인터넷” 단어를 발화하였을 때 MFCC 기법으로 추출된 특징이다. Fig. 5는 64개의 단어들에 대해 Augmentation 기법인 Mixup을 적용한 스펙트럼이다. 그러나 Fig. 4의 스펙트럼의 본질적인 특징들은 오히려 손실된 것처럼 보인다. 충분한 데이터를 학습하지 못 할 때에는 오히려 데이터 고유의 특징을 학습하지 못 할 수 있기 때문에, 충분한 수량으로 구축된 데이터 셋과 적절하게 Mixup을 선택적으로 사용하는 것이 필요하다.

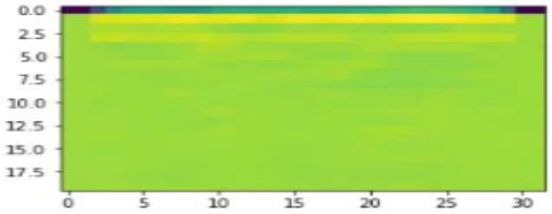

Fig. 4. MFCC of the word “Internet”

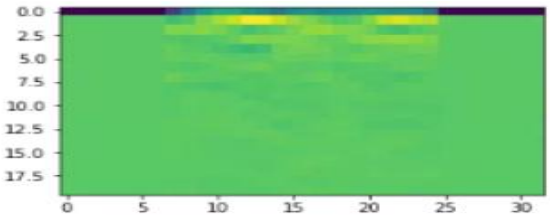

Fig. 5. Mixup Spectrum with 64 words (e.g., Internet, E-mail, etc.)

ResNet은 학습 중 모델이 과적합되지 않도록 학습률(Learning rate)을 고원(Plateau) 부분에서 조정하였으며, 확률적 경사 하강법(SGD)을 사용 시, 최적화 모델은 모멘텀 (Momentum)을 적용하였다. CNN1은 학습률을 단계적으로 줄여나갔으며, 속도를 계산하여 학습의 갱신강도를 적응적으로 조정해나가게 했다.

Table 3. Speech Command Recognition Results of “Healthcare”

Model	Pure	Mixup
CNN1	25.0%	27.0%
ResNet8	77.0%	43.7%
ResNet15	**89.5%**	52.0%

Table 4. Speech Command Recognition Results of “Daily Lives”

Model	Pure	Mixup
CNN1	14.0%	8.0%
ResNet8	**82.0%**	48.0%
ResNet15	80.0%	34.0%

Table 3과 Table 4의 Pure는 음성데이터에서 어떤 처리도 하지 않은 채, MFCC자질만 추출한 데이터를 학습한 모델의 성능이며, Mixup은 Pure실험에서 증강기법인 Mixup을 적용하여 학습한 모델의 성능이다. CNN1은 Chen et al.[9]의 CNN의 전형적인 기법인 cnn-trad-fpool3모델이며, 두 개의 ResNet기반 모델들은 Tang et al.[11]에서 Warden [10]의 데이터 셋 기반으로 성능이 검증된 ResNet8과 ResNet15를 사용하였다.

건강 도메인에서 ResNet가 89.5%로 가장 높은 성능을 보였으나, CNN1을 사용한 경우 상당히 낮은 정확도를 보

였다. ResNet8의 파라미터 개수가 117K로 CNN1대비 약 70K 더 적음에도 불구하고 더 높은 성능을 보였다.

Zhang et al.[6]에서는 모델의 깊이가 더 깊을수록 Mixup을 사용할 때, 성능이 더 향상됨을 보였으나 본 실험의 데이터를 사용하였을 때는 CNN1을 제외한 ResNet 기반 모델의 정확도가 크게 떨어지는 것을 확인하였다.

반면에, 일상생활 도메인에서는 ResNet8이 가장 높은 성능을 보였다. Choi et al.[14]에서는 모델의 깊이가 깊어질수록 더 좋은 성능을 개선된 연구가 있다. 그럼에도 불구하고 경량화된 모델이 훨씬 높은 성능을 내는 상황이 존재 할 때도 있다. 특히, 건강 도메인의 데이터를 학습을 한 ResNet15 모델의 성능이 낮지 않음을 고려할 때, 일상생활 도메인의 데이터를 학습하면 비교적 모델의 학습 가능한 파라미터의 크기가 더 큰 ResNet15에서는 각 데이터를 학습한 컨볼루션 레이어들의 가중치에서 희박한 부분들이 더 많이 존재하게 된다. ResNet계열의 알고리즘에 비해 베이스라인으로 활용된 CNN1은 현저히 낮은 수준의 정확도를 보였다.

IV. Discussion

자동 구축 데이터들의 노이즈: 자동 구축된 데이터들을 실제 사람이 들어보면, 각 데이터 샘플파일별로 서로 상이하면서, 구분이 되지 않는 노이즈가 많이 존재하여, 실험에서 성능의 향상이 매우 제한적인 것으로 추정된다. 즉, 학습데이터에 포함된 다양한 유형의 노이즈가 성능에 많은 영향을 끼치고 있는 것으로 보인다. 반면, 다양한 유형의 노이즈가 존재하기에 대량의 데이터로 학습된 모델의 평가용도로 사용하면 유용할 것 같다.

자동 구축의 한계: 건강 및 일상생활 도메인에서 4000여 개의 단어 별로 약 2만여 개의 음성 샘플 데이터가 수집되었다. 하지만, 그 수는 일련의 정제과정을 거치면서 큰 폭으로 감소되었는데, 실제 학습데이터로 채택한 수는 최초 수집데이터의 2.5~2.7% 수준이다. 한국어 YouTube의 자막은 실제 단어 수준 발화에 있어 싱크 매칭 비율이 매우 낮음을 추정할 수 있다. 또한 Google Cloud의 STT API를 이용하여 음성 데이터 정제 시에 완벽히 매칭되지 않으면 데이터 셋에서 제거되어 데이터 구축 면에서 효율성이 좋지 않았다. 따라서, 원하는 개수의 학습데이터를 자동으로 구축하기 위해서는 더 많은 수의 동영상 컨텐츠를 대상으로 추출작업을 진행해야 할 수 있다.

특징증가의 문제점: 이미 Mixup기법 자체는 기존 연구 Zhang et al.[6]에서 성능 효율성이 검증되어 있음에도 불구하고, 데이터에 따라 본래의 정보를 손실하게 되는 경향이 있음이 확인되었다. 본 연구의 실험에서 Mixup의 적용이 그리 효과적이지 못했다. 따라서, 노이즈가 많이 섞여 있는 음성 명령어 데이터에 맞은 데이터 증강기법을 모색하고 추가적인 실험을 할 필요가 있다.

학습 데이터 자동 구축의 가능성: 자동으로 구축된 데이터가 약 80%대의 좋은 성능을 보임으로 인해 실질적으로 여러 음성인식 분야에 폭넓게 응용이 가능하다.

Lakomkin et al.[1]에서는 WSJ(Wall Street Journal) 데이터와 TED데이터, YouTube데이터를 이용하여 WER(Word Error Rate)과 CER(Character Error Rate)을 측정 했는데, WSJ 데이터와 YouTube 데이터를 섞거나, TED 데이터와 YouTube 데이터를 섞을 때 각각의 Error Rate가 가장 낮게 나왔다. 즉, 기존 데이터와 자동 구축한 데이터를 합하여 성능향상을 보였다. 현재 구글 STT를 이용한 정제 시 의외로 사람이 들었을 때 취할 수 있는 음성 샘플도 탈락하는 현상이 다소 발생한다. 따라서 보다 개선된 음성데이터 비교·정제기법을 사용하여 정제 시 제외된 데이터 중 일부를 추가적으로 반영한다면 현재보다 훨씬 많은 데이터를 자동으로 구축할 수 있을 것으로 기대된다.

V. Conclusions

본 논문에서는 음성인식 데이터 부족 문제를 해결하기 위한 방안으로 제시된 자동 데이터 구축 기술을 사용하여 직접 음성데이터를 구축하고 수정된 ResNet모델을 이용하여 자동 구축된 데이터의 활용가능성을 보였다.

구축된 데이터는 건강과 일상생활 도메인에서 각각 15개씩 명령어를 선정하여 수집하였다. 또한, 제안된 데이터 정제 기법을 통해 최종적으로 550개와 552개의 음성 명령어 발화를 정제하여 데이터 셋을 구축하였다.

전형적인 CNN 기반 모델을 사용할 때 현저히 낮은 정확도를 보였지만, ResNet 기반 모델을 사용하였을 때, 건강 도메인에서는 89.5%, 일상생활 도메인에서는 82.0%의 정확도를 증명하였다. 정제과정에서 데이터의 양이 많이 감소되어 충분한 데이터를 확보하지 못함에도 불구하고 낮지 않은 성능을 확인하였다.

따라서 웹을 출처로 하여 자동으로 음성인식 학습 데이터를 구축하여 활용하는 것은 매우 유용하다는 것을 알 수

있다. 제안기법은 음성 명령어 인식이 요구되는 다양한 도메인의 비즈니스 (예, AI스피커, 식당, 자동차, 항공기 [15] 등)에서 추가적인 학습데이터의 확장이 필요한 곳에 적용되어 저비용으로 학습데이터를 구축할 수 있을 것이다. 현재 구글의 STT를 이용한 음성 정제기법만을 사용 중인데, 향후에 보다 고도화된 정제기법이 적용된다면, 보다 많은 수의 유효 데이터를 확보 할 수 있으면, 이는 모델의 정확도 개선으로 이어질 수 있다. 향후 음성데이터에 대해 적대적 사례 (adversarial example)를 역으로 활용하거나 [16], 강화학습 [17]을 도입하여 추가적인 성능향상이 가능한지 검증 해볼 필요가 있다.

ACKNOWLEDGEMENT

This research was financially supported by the Ministry of Trade, Industry and Energy (MOTIE) and Korea Institute for Advancement of Technology (KIAT) through the National Innovation Cluster R&D program (P0006704).

REFERENCES

[1] E. Lakomkin, S. Magg, C. Weber, and S. Wermter, “KT-Speech-Crawler: Automatic Dataset Construction for Speech Recognition from YouTube Videos,” arXiv:1903.00216 , 2019.

[2] Zeroth project, Available at https://github.com/goodatlas/zeroth

[3] KSS data set, Available at https://www.kaggle.com/bryanpark/korean-single-speaker-speech-dataset

[4] J. Kaewprateep and S. Prom-On, “Evaluation of small-scale deep learning architectures in Thai speech recognition,” 1st Int. ECTI North. Sect. Conf. Electr. Electron. Comput. Telecommun. Eng. ECTI-NCON 2018, pp. 60–64, 2018.

[5] Y. Choi and B. Lee, “Pansori: ASR Corpus Generation from Open Online Video Contents,” IEEE Seoul Sect. Student Pap. Contest, pp. 117–121, 2018.

[6] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “MixUp: Beyond empirical risk minimization,” 6th Int. Conf. Learn. Represent. ICLR 2018 - Conf. Track Proc., pp. 1–13, 2018.

[7] T. N. Sainath and C. Parada, “Convolutional Neural Networks for Small-footprint Keyword Spotting,” Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, pp. 1478–1482, 2015.

[8] R. Tang and J. Lin, “Honk: A PyTorch Reimplementation of Convolutional Neural Networks for Keyword Spotting,” arXiv:1710.06554, 2017.

[9] G. Chen, C. Parada, and G. Heigold, “Small-footprint keyword spotting using deep neural networks,” ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., pp. 4087–4091, 2014.

[10] P. Warden, “Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition,” arXiv:1804.03209, 2018.

[11] R. Tang and J. Lin, “Deep residual learning for small-footprint keyword spotting,” ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., vol. 2018-April, pp. 5484–5488, 2018.

[12] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2016-December, pp. 770–778, 2016.

[13] K. He, X. Zhang, S. Ren, and J. Sun, “Identity mappings in deep residual networks,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9908 LNCS, pp. 630–645, 2016.

[14] S. Choi et al., “Temporal convolution for real-time keyword spotting on mobile devices,” Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, vol. 2019-September, pp. 3372–3376, 2019.

[15] D. Oneață and H. Cucu, “Kite: Automatic Speech Recognition for Unmanned Aerial Vehicles,” Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, vol. 2019-September, pp. 2998–3002, 2019.

[16] T. Rajapakshe, R. Rana, S. Latif, S. Khalifa, and B. W. Schuller, “Pre-training in Deep Reinforcement Learning for Automatic Speech Recognition,” arXiv:1910.11256, 2019.

[17] J. Vadillo and R. Santana, “Universal adversarial examples in speech command classification,” arXiv:1911.10182, 2019.

Authors

Yeonsoo Lim is currently an undergradute student of the Department of Computer Engineering at Kumoh National Institute of Technology. His research interests include NLP, Speech Recognition, and Deep Learning.

Deokjin Seo is currently an undergradute student of the Department of Computer Engineering at Kumoh National Institute of Technology. His research interests include Speech Recognition, Deep Learning, and Crawling.

Jeong-sik Park received his B.E. degree in Computer Science from Ajou University, South Korea in 2001 and his M.E. and Ph.D. degree in Computer Science from KAIST (Korea Advanced Institute of Science and Technology) in 2003 and 2010, respectively. Dr. Park was a Post-Doc. researcher in the Computer Science Department, KAIST ('10~'11). And he joined the faculty of the Department of Intelligent Robot Engineering, Mokwon University, as an assistant professor ('12~'13), and the faculty of the Department of Information and Communication Engineering, Yeungnam University, as an assistant professor ('14~'17). He is now an associate professor in the Department of English Linguistics & Language Technology, Hankuk University of Foreign Studies. His research interests include speech signal processing, speech recognition, and voice interface for human-computer interaction.

Yuchul Jung received his B.E. degree in Computer Science from Ajou University, South Korea in 2001 and his M.E. in Information & Communication Engineering and Ph.D. degree in Computer Science from KAIST (Korea Advanced Institute of Science and Technology) in 2005 and 2011, respectively. Dr. Jung joined the faculty of the Department of Computer Engineering at Kumoh National Institute of Technology (KIT), Gumi, in 2017. Prior to joining KIT, he worked a senior researcher at Korea Institute of Science and Technology Information (KISTI) ('13~'17) and Electronics and Telecommunications Research Institute (ETRI) ('09~'13), Daejeon, South Korea. His research interests include machine learning based NLP (text mining, sentiment analysis, automatic knowledge base construction, etc.), Korean speech recognition, and Medicine 2.0.