Journal of KIIT. Vol. 16, No. 11, pp. 11-21, Nov. 30, 2018. pISSN 1598-8619, eISSN 2093-7571 11
http://dx.doi.org/10.14801/jkiit.2018.16.11.11

CNN을 이용한 음성 데이터 성별 및 연령 분류 기술 연구

박대서*[1], 방준일*[2], 김화종**, 고영준***

A Study on the Gender and Age Classification of Speech Data Using CNN

Dae-Seo Park*[1], Joon-Il Bang*[2], Hwa-Jong Kim**, and Young-Jun Ko***

2016년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No: R-20160906-004163, 빅데이터 자동 태깅 및 태그 기반 DaaS 시스템 개발)
2018년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No: C1014236-01-01, 지능형 디바이스 및 게이트웨이/허브를 위한 사용자 친화적 지능형 UI/UX 기술 구현)

요 약

본 논문에서는 사람을 대신하여 분류, 예측 하는 딥러닝 기술을 활용하여 목소리를 통해 남녀노소를 분류하는 연구를 수행한다. 연구과정은 기존 신경망 기반의 사운드 분류 연구를 살펴보고 목소리 분류를 위한 개선된 신경망을 제안한다. 기존 연구에서는 도시 데이터를 이용해 사운드를 분류하는 연구를 진행하였으나, 얕은 신경망으로 인한 성능 저하가 나타났으며 다른 소리 데이터에 대해서도 좋은 성능을 보이지 못했다. 이에 본 논문에서는 목소리 데이터를 전처리하여 특징값을 추출한 뒤 추출된 특징값을 기존 사운드 분류 신경망과 제안하는 신경망에 입력하여 목소리를 분류하고 두 신경망의 분류 성능을 비교 평가한다. 본 논문의 신경망은 망을 더 깊고 넓게 구성함으로써 보다 개선된 딥러닝 학습이 이루어지도록 하였다. 성능 결과로는 기존 연구와 본 연구의 신경망에서 각각 84.8%, 91.4%로 제안하는 신경망에서 약 6% 더 높은 정확도를 보였다.

Abstract

Research is carried out to categorize voices using Deep Learning technology. The study examines neural network-based sound classification studies and suggests improved neural networks for voice classification. Related studies studied urban data classification. However, related studies showed poor performance in shallow neural network. Therefore, in this paper the first preprocess voice data and extract feature value. Next, Categorize the voice by entering the feature value into previous sound classification network and proposed neural network. Finally, compare and evaluate classification performance of the two neural networks. The neural network of this paper is organized deeper and wider so that learning is better done. Performance results showed that 84.8 percent of related studies neural networks and 91.4 percent of the proposed neural networks. The proposed neural network was about 6 percent high.

Keywords
voice, classification, neural network, deep learning, classification, CNN,

* 강원대학교 컴퓨터정보통신공학과
- ORCID[1]: https://orcid.org/0000-0001-7421-5926
- ORCID[2]: https://orcid.org/0000-0003-0582-1572
** 강원대학교 컴퓨터정보통신공학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-3822-390X
*** ㈜ 모다정보통신 IoT 개발단장
- ORCID: https://orcid.org/0000-0001-9695-6119

· Received: Sep. 07, 2018, Revised: Oct. 05, 2018, Accepted: Oct. 08, 2018
· Corresponding Author: Hwa-Jong Kim
Dept. of Computers and Communications Enginieering Kangwon National Univ., Gangwondaehak-gil 1, ChunCheon-si, Gangwon-do, Korea
Tel.: +82-33-250-6323, Email: hjkim3@gmail.com

Ⅰ. 서 론

1.1 연구 목적 및 동기

사람은 목소리를 통해 감정, 의도 등을 전달하여 상대방과 의사소통 한다. 목소리의 높낮이, 속도, 어휘 등을 기준으로 보지 않고도 상대방을 특정하고 기분을 파악하고 목적이 무엇인지 알아챌 수 있다. 단순히 보았을 때 하나의 소리에 불과한 사람의 음성을 통해서 다양한 정보를 도출해 낼 수 있는 것이다. 이렇게 많은 정보를 담고 있는 음성 데이터를 딥러닝을 통해 분석함으로써 텍스트 변환, 감정분석, 사용자 분류, 음성검색 등에 활용할 수 있다[1].

최근 음성 분석 기술은 딥러닝과 결합하여 과거에 비해 큰 향상을 이뤄내고 있으며 인공지능 스피커, 스마트폰, 검색엔진 등에 탑재되어 많은 사람들이 손쉽게 이용 가능하다. 현재 서비스되고 있는 음성 기술은 목소리를 입력받아 특징을 추출하고 언어모델, 어휘사전, 발음규칙 등을 고려하여 사용자의 목적을 분석하고 그에 맞는 검색, 추천 서비스를 제공한다[2]. 음성 분석 기술은 과거에 비해 큰 성과를 내고 있지만, 주변의 소음 여부 등에 따라서 인식률의 감소가 불가피한 상황이다. 또한, 같은 단어, 문맥일지라도 사람마다 발성의 차이가 있기 때문에 지속적인 연구와 발전이 필요한 상황이다. 분석과정에서 고려되지 않은 단어, 어휘, 지역별 사투리 등은 음성 인식률을 낮추는 요인이 된다[2].

국내외의 기업에서 음성인식 정확도 향상을 위한 연구가 지속되고 있으며 인공지능 기술로 아이비엠 트루노스 구글 웨이브넷, 아마존 알렉사 등이 대표적이다[3]-[5]. 최근 주된 음성 분석 기술은 인공지능 기반 음성비서 기술과 인공지능 스피커 기술로 국내에서도 SKT, 카카오, 네이버 등에서 자사의 음성 인식기능을 탑재한 인공지능 스피커를 출시하였다. 사실상 인공지능 스피커는 서비스를 위한 도구이기 때문에 핵심은 음성 비서 기술이라 할 수 있으며, 스마트폰, 스마트워치, 로봇, 자동차 등에 다양하게 탑재되고 있다[5]. 하지만, 이렇게 국제적인 연구와 서비스화에도 불구하고 음성인식 서비스 이용률은 낮은 상태이며 서비스 확대와 사용자 유입을 위해서는 무엇보다 성능개선이 절실한 상황이다.

따라서 본 논문에서는 음성 분석 기술의 성능 개선에 초점을 맞추었으며 사람의 음성 데이터를 이용해 모델을 학습하고 주어진 음성에 대해 남녀노소 클래스를 분류하는 기술을 연구한다. 연구는 선행연구에서 소개 된 음성 분류 신경망과 본 논문에서 제안하는 신경망을 비교한다. 분석 과정은 수집된 음성 데이터로부터 특징값을 추출해 벡터화[6]하여 신경망 모델을 학습시키고 테스트 데이터에 대한 클래스 분류 성능을 도출하여 두 모델의 성능 평가와 최종 결론을 도출한다. 신경망은 두 모델 모두 합성곱 신경망[7]으로 구성되며 분석 프로그래밍은 파이썬을 활용해 전처리, 분석, 시각화 등을 처리한다.

1.2 연구 필요성

사용자의 음성을 입력함으로써 기존에 설문조사처럼 요청하던 나이, 성별 등의 개인 정보를 간단하고 빠르게 처리할 수 있다면 사용자와 서비스 제공자 측면에서 큰 이점이 있을 것으로 생각된다[8][9]. 사용자의 음성을 통해 기본적인 정보(성별, 연령대)를 추출하고 그 결과를 사용자에게 확인받는다면 요청 단계를 축소하고 사용자의 편의성 향상에도 기여할 수 있다. 이에 본 논문에서는 음성 분석을 통해 사용자를 분류하는 연구를 진행하여 기존 연구[10]보다 개선된 신경망 모델을 제안한다.

1.3 연구 방법

연구는 기존 연구[10]에서 제시한 신경망 모델과 본 논문에서 제시하는 모델을 비교 분석한다. 우선, 기존 연구의 데이터처리, 모델 구성, 분석결과를 통해 논문의 내용을 파악하고 정리한다. 다음으로 본 연구의 차별성을 제시하고 음성 데이터 수집부터 결과도출까지 전체적인 연구를 진행하여 기존 연구와의 성능차이를 비교 평가한다. 데이터셋은 기존 연구에서 활용한 데이터셋이 아닌 본 논문의 목적에 맞는 음성 데이터를 활용하며 모델 성능 비교를 위해 모델 학습 단계부터 최종 결론 도출까지 두 개의 신경망을 동일한 조건(데이터 구성, 학습과정) 아래에서 학습한다.

표 1. 분석 컴퓨터 사양
Table 1. Analysis computer specifications

Composition	Spec.
OS	ubuntu 16.04 LTS - 64bit
Memory	62GB
CPU	Intel Core i7-6700K CPU @ 4.00GHz * 8
GPU	TITAN X
Disk	1TB

표 1은 본 연구에서 신경망 모델 학습을 위해 사용한 컴퓨터 환경이다.

전체적인 분석과 신경망 구축은 파이썬, 케라스를 사용하고 모델 학습에는 그래픽카드를 사용한다. 음성 데이터 처리 알고리즘과 평가함수, 시각화 도구를 통해 분석 및 결과 평가를 진행한다.

II. 관련 연구

2.1 선행연구

본 논문에서는 사람의 음성을 분석하여 연령대를 구분하는 연구를 수행할 것이기 때문에 기존 연구의 소리 분류 신경망을 사람 음성 분류에 적용하였을 때 성능을 분석하고 개선된 신경망 모델을 제안한다.

신경망을 통한 소리 분류 연구는 주로 도시 내에서 발생하는 자동차, 공사장, 비행기 등의 소음을 분류하거나 가정 내에서 발생하는 헤어드라이기, 전자레인지 등과 같이 서로 다른 카테고리에서 발생하는 소리를 분류하는 연구가 이루어지고 있다[10][11]. 하지만, 소리 분류 연구 중 음성 분류는 수많은 사람들의 음성정보를 확보해야하는 어려움이 있고, 데이터 보유자 또한 개인정보 등의 이유로 공개하지 않기 때문에 데이터는 더욱 찾아보기 힘들다. 따라서 본 연구에서는 대체재로써 기존 연구 중 도시 소음 데이터 분류 연구의 특징, 성능, 문제점과 개선사항을 살펴보고 화자 분류를 위한 신경망 모델을 제안한다.

[10]에서는 8723개의 도시 데이터(Urban-Sound8k)를 사용하였으며 클래스는 에이컨(Air conditioner), 자동차(Car horn), 어린이(Children playing), 강아지(Dog bark), 드릴(Drilling), 엔진(Engine idling), 총(Gun shot), 착암기(Jackhammer), 사이렌(Siren), 노래(Street music) 총 10개로 구성되어 있으며 각 클래스 별 데이터 수는 표 2와 같다.

표 2. Urban-Sound8k 클래스
Table 2. Urban-sound8k class

Class	Number
air conditioner	1000
car horn	429
children playing	1000
dog bark	1000
drilling	1000
engineidling	1000
gun shot	374
jackhammer	1000
siren	929
street music	1000

표 2의 데이터로부터 로그 스케일 멜-스펙트로그램(Mel-Spectrogram)[6]을 특징값으로 추출하여 신경망 학습에 활용하였다. 분류를 위한 신경망 구성은 단일 합성곱 신경망(CNN)을 여러 층으로 구성한 다층 합성곱 신경망 모델을 제안하였다.

신경망의 구성은 그림 1과 같이 4개의 합성곱 신경망계층으로 구성하였고 계층별로 맥스풀링과 드롭아웃, 활성 함수로 렐루(Relu)를 구성하였으며 손실률 계산을 위해 크로스-엔트로피, 아담 옵티마이저로 신경망을 구성하였다[12][13].

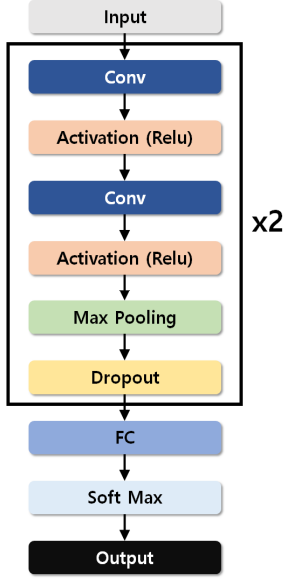


그림 1. CNN 구성 [2]
Fig. 1. CNN Configuration [2]

표 3. 클래스 별 성능
Table 3. Class performance

Class	Accuracy
air conditioner	0.87
car horn	0.88
children playing	0.73
dog bark	0.81
drilling	0.81
engineidling	0.86
gun shot	0.93
jackhammer	0.89
siren	0.88
street music	0.77

신경망 학습은 32 크기의 미니배치, 에포크 50으로 고정하여 진행하였으며 제안하는 신경망에서 평균 정확도는 0.83으로 [10]에서 비교분석한 다른 모델보다 높은 수준을 보였다. 또한, 클래스 별 성능은 표 3과 같다. 표 3에서 어린이(Children playing)와 노래(Street music)는 상대적으로 낮은 성능을 보였지만, 전체적으로 준수한 성능을 보이고 있다. 그림 2를 통해서 클래스별로 올바르게 분류한 수와 잘못 분류했을 경우 어떤 클래스로 잘못 분류했는지 확인 할 수 있다. 주로 어린이와 노래에서 상호간에 혼동해서 분류된 경우가 눈에 띈다.

[10]의 분석 결과는 준수한 예측 결과를 보이고 있지만, 100%의 성능을 기대했을 때 다소 낮다고 평가 할 수 있으며 여러 요인에 따라서 더 높은 성능을 낼 수 있을 것으로 판단된다. 또한, 신경망 구성에 있어서도 더 깊고 넓은 형태로 구성함으로써 성능 개선과 다른 결과를 도출 할 수 있을 것으로 기대된다. 마지막으로, 분석 결과는 훈련셋과 테스트셋에서 살펴본 것의 성능 차이가 날 수 있기 때문에 모델 성능의 정확한 판단을 위해서는 테스트셋에 대한 성능 도출과 신경망 학습 단계에서 손실률 등이 적합하게 갱신되는지도 살펴 볼 필요가 있다. 본 연구에서는 소리 분류를 위한 공통의 연구를 진행하지만, 데이터를 도시데이터에서 사람 음성 데이터로 변환하였을 때 [10]에서 제안한 신경망이 좋은 성능을 보이는지 연구하고, 최종적으로 더 나은 신경망과 요인들을 구성하여 높은 성능 결과를 도출하도록 한다.

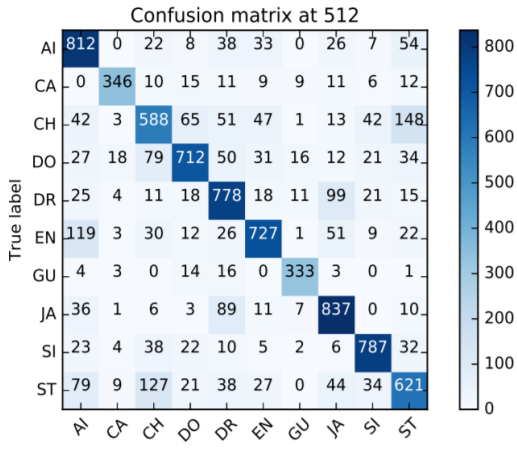


그림 2. 분석 결과 혼돈 매트릭스
Fig. 2. Confusion matrix of result

2.2 본 연구의 차별성

선행 연구의 도시 데이터 분류 기술에서는 합성곱 신경망(CNN)을 얕은 신경망으로 구성하여 음성 분류를 수행하였으며 10개의 음성에 대해 60%의 성능을 보였다. 최신 신경망 기술을 도입한 것을 감안하였을 때 높은 성능이라 판단하기 어렵다. 하지만, 본 연구의 목적은 사람의 음성을 학습하여 남녀노소를 구분하는 연구이기 때문에 선행 연구의 신경망을 그대로 활용하고 화자 데이터를 학습하였을 때 성능이 어떻게 나타나는지 파악한 후에 연구 자료로 활용할 것이다.

본 연구에서는 도시 데이터가 아닌 사람의 음성을 통해 화자를 5개 클래스로 분류하는 연구를 수행할 것이며 이 연구 과정에 선행 연구에서 활용한 신경망과 본 논문에서 제안하는 신경망의 성능차이를 비교한다.

핵심 차별점으로 선행 연구는 각 층에 1개의 컨볼루션 필터를 구성하여 얕은 신경망을 구성하였지만 본 연구에서는 신경망을 깊게 구성하고 각 층에 3개의 컨볼루션 필터를 구성한다. 한 개의 층에 여러 개의 컨볼루션 필터를 구성함으로써 단일 필터 일 때보다 정확도를 올리고 컴퓨터 작업량을 줄일 수 있다[14][15]. 망이 깊을수록 레이어가 넓을수록 높은 성능을 보이는 아이디어에 준하여 신경망을 구성한다.

III. 음성 분류

3.1 연구 프로세스 및 데이터 수집

본 논문의 연구 과정은 그림 3의 단계별로 데이터 수집, 1차 전처리, 2차 전처리, 모델 구성(신경망), 모델 학습, 결과 평가 순으로 진행한다.

화자 분류를 위한 음성 데이터는 외국 사이트의 무료 음성 데이터와 유튜브 영상으로부터 추출하여 활용하였다. 데이터의 클래스는 여성, 남성, 어린이, 노인 여성, 노인 남성 총 5개 클래스로 분류하여 수집하였다.

3.2 데이터 전처리

전처리 과정은 자체적으로 2단계로 구성하였는데 1차 전처리에서는 음성의 공백 제거와 음성 데이터 간에 길이 차이를 완화하기 위한 음성 커팅 작업을 진행한다. 커팅 작업 시 3초 이상 음성이 나타나지 않을 때를 기준으로 하여 커팅 작업을 수행한다. 여기까지가 1차 전처리 단계이며 2차 전처리는 모델에 입력하기 전에 데이터를 균일한 크기로 벡터화하여 모델에 입력할 수 있도록 처리한다. 2차 전처리는 1차 전처리에서 커팅 된 특정 음성 데이터 파일을 로드하고 일정 구간마다 음성 정보를 추출하여 벡터화하는 작업을 수행한다[16]. 음성데이터의 벡터화에는 멜-스펙트로그램을 적용한다[6].

사람은 소리를 비선형적으로 받아들이며 멜 곡선(Mel-Curve)은 사람의 청각 특성을 반영한다. 사람은 저주파의 경우 작은 변화에도 민감하게 반응하지만, 고주파일수록 잘 구별하지 못한다(비선형성). 이러한 원리를 적용하는 멜-스펙트로그램은 주파수 성분을 멜 곡선에 따라 압축(Mel-scale)하고 로그를 취해 얻어 낼 수 있다. 음성 데이터를 시각화 할 경우 그림 4처럼 확연히 다른 형태로 나타난다. 전처리 시 원본 음성을 지정된 크기만큼(Window) 여러 구간으로 분할하여 각각의 구간에 대해 멜-스펙트로그램을 추출해 데이터를 벡터화 한다. 스펙트로그램 추출을 위해서는 음성(y), 주파수 축 범위, 음성(y)의 초당 샘플링 수, 특징 추출을 위한 프레임 간 샘플 수를 설정하여 진행한다.

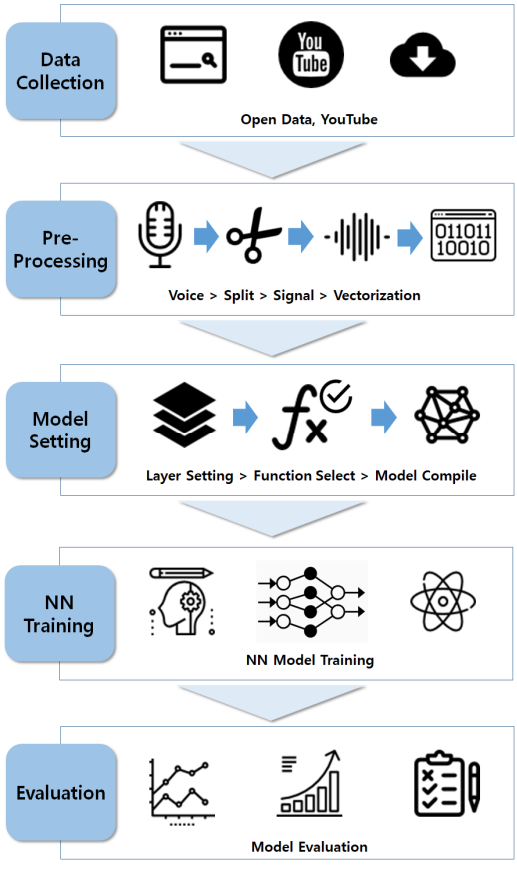


그림 3. 분석 프로세스
Fig. 3. Analysis process

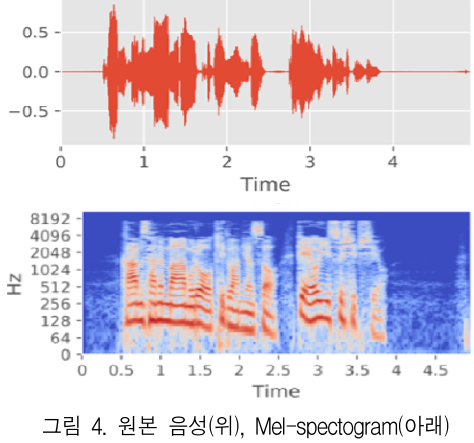


그림 4. 원본 음성(위), Mel-spectogram(아래)
Fig. 4. Original voice(up), mel-spectrogram(down)

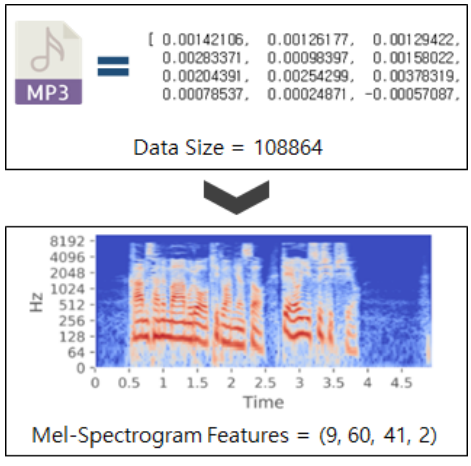


그림 5. 원본 음성의 특징 추출
Fig. 5. Feature extraction of original voice

본 논문에서는 1차 전처리 된 음성과 주파수 축 범위 60, 구간 크기 20480(512*40), 음성의 초당 샘플링 수 22050, 프레임 간 샘플 수 512로 설정하여 벡터정보를 추출하고 로그를 취해 최종 데이터 벡터를 추출한다.

2차 전처리를 통해 그림 5처럼 원본 데이터로부터 멜-스펙트로그램을 추출하여 분석용 데이터를 생성한다. 데이터는 (N, 주파수 축 범위, 프레임 수, 채널 수)의 형태로 재배열하여 (N, 60, 41, 2)으로 특징 값을 벡터화한다. 특징값을 벡터화 한 후 0 ~ 4의 숫자로 정해진 레이블 값을 원핫인코딩하여 딥러닝 연산 시 크로스 엔트로피와 소프트맥스를 적용한다.

분류는 남, 여 2개 클래스 분류와 남, 여, 어린이, 노인 남, 노인 여 5개 클래스에 대해 실시하며 2개 클래스 데이터 크기 N은 9723개로 훈련 데이터 7414개, 검증 데이터 1309개, 테스트 데이터 1000개로 나누어서 학습을 진행한다. 5개 클래스 데이터는 9284개로 훈련 데이터 7041개, 검증 데이터 1243개, 테스트 데이터 1000개로 구성하여 연구를 수행하였다. 데이터 배정에 정립된 기준은 없으며 약 75%를 훈련, 15%를 검증 10%를 테스트데이터로 배정하여 학습 및 평가에 활용하였다. 5개 클래스의 수가 적은 이유는 2개 클래스의 경우 국내외 무료 데이터가 상당히 많이 공개되어 있었으며 별도의 녹음, 추출 없이 활용 가능한 데이터가 많았으나 어린이, 노인의 경우 상당히 수가 적었으며 동영상등으로부터 별도 추출하여 활용하였기 때문에 남녀 데이터에 비해 상대적으로 적은 상황이 발생하였다. 따라서 클래스별 데이터 수의 격차가 크지 않도록 다른 클래스와 비율을 맞추어 데이터를 구성하였다.

3.2 신경망 모델 구성

앞에서 언급한 것과 같이 깊고 넓은 망을 구축하기 위하여 파이썬 케라스를 활용하였으며 그림 6과 같이 신경망을 구성하였다[17].

합성곱 신경망을 적용하는 점에서는 선행 연구와 동일하지만 망의 깊이와 하나의 층에 3개의 합성곱 신경망을 적용한 점에서 차이가 있으며 활성함수를 리키렐루(LeakyReLU)로 구성하였다[18].

리키렐루는 각 뉴런의 출력값이 0보다 높으면 그대로 두고, 0보다 낮으면 정해진 숫자를 곱하는 방식의 함수이다. 렐루(ReLU)의 경우 0보다 작을 때 경사가 사라지는 문제가 있지만 리키렐루의 경우 0보다 작을 때도 미분을 적용할 수 있다. 최적화 함수는 아담 옵티마이저, 컨벌루션 필터는 (1,1), (3,3), (5, 5) 3개를 혼합하여 적용하였다.

모델구성은 그림 6과 같이 입력 직후 단일 신경망 계층을 통과시켜 그 출력을 병렬로 구성된 3개의 신경망 계층으로 동시에 입력한다. 이후 3개의 계층으로부터 나온 출력을 하나로 병합해 동일한 과정을 2회 반복한다. 끝으로 2개의 단일 신경망 계층을 통과시키고 소프트맥스 함수를 적용해 분류 결과를 출력한다.

3.3 모델 학습

앞서 2차 전처리 단계에서 설명한대로 클래스 수별로 목적에 따른 데이터셋을 구성한다. 표 4에 훈련, 검증, 테스트 데이터셋 수를 작성하였다.

표 4. 클래스 별 데이터셋 구성
Table 4. Dataset by class

Class Number	Train	Valid	Test	Total
2	7,414	1,309	1,000	9,723
5	7,041	1,243	1,000	9,284

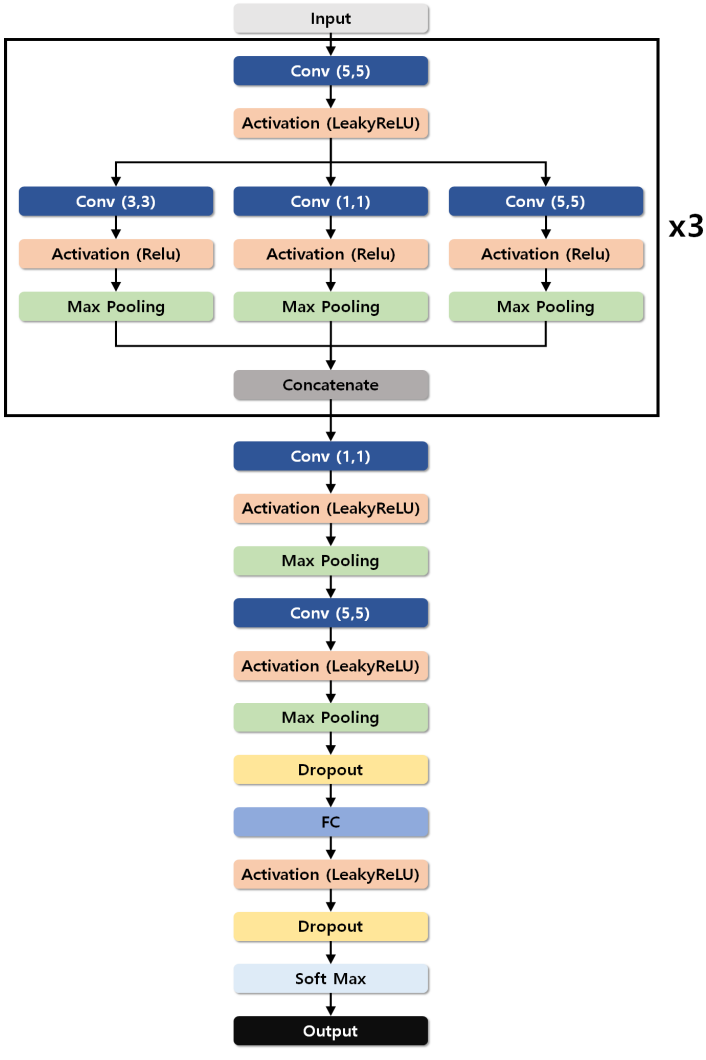


그림 6. 본 연구의 제안 신경망
Fig. 6. Proposed neural network of this paper

훈련, 검증, 테스트 데이터로 각각 나누기 위하여 전체 범위에서 무작위로 추출하여 각각 검증 데이터와 테스트 데이터로 나눈다. 데이터 셋을 나누는 과정에서 사람이 개입할 경우 모델 훈련 과정과 성능 결과의 신뢰성이 떨어질 수 있고 클래스별로 잘 섞이지 않을 경우 데이터의 순서에 따라 과적합과 성능저하 현상이 나타날 수 있기 때문에 무작위로 지정된 크기만큼 각 데이터 셋에 배정한다[19][20].

모델 학습은 두 개의 신경망 모델을 동일한 컴퓨터 환경에서 진행하였으며 각 학습 회차(Epoch)마다 훈련 셋과 검증 셋의 손실값과 정확도를 계산하여 정상적으로 학습이 이루어지는지 확인하였다. 학습은 총 30회에 걸쳐 [10]과 제안 모델 대해 동일하게 진행하였으며 2개, 5개 클래스 분류를 각각 학습 하였다. 모델 학습 후 손실함수 그래프는 표 5에 기록하였으며 신경망 모델별, 클래스 수 별 학습 그래프를 나타낸다.

학습된 모델은 학습 과정에서 나타난 손실값과 정확도를 기준으로 정상적인 학습이 이루어졌는지와 어느 정도의 성능이 나타나는지 판단 할 수 있으며 최종적으로 테스트 셋을 입력하여 올바르게 분류를 해 내는지 확인함으로써 모델을 평가할 수 있다[19][21].

표 5 훈련, 검증 성능
Table 5. Training, validation performance

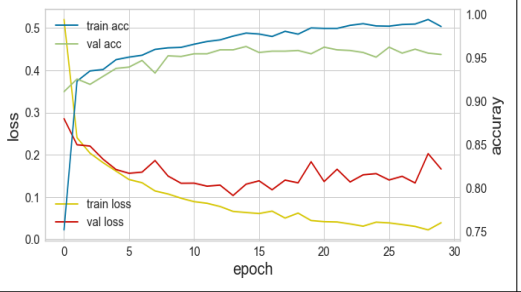
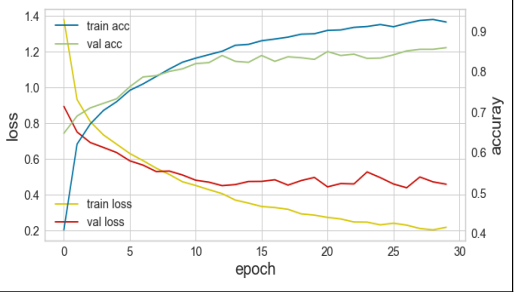
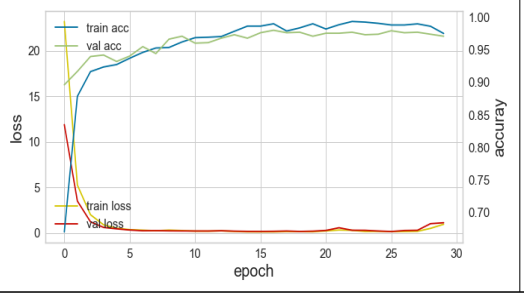
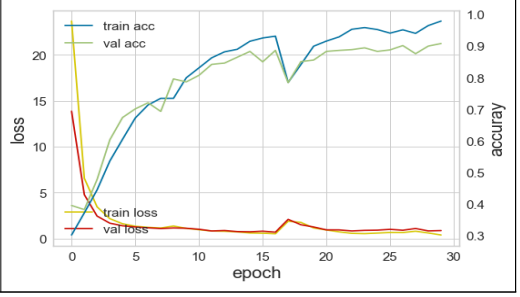
	2 Class (Man, Woman)	5 Class (Man, Woman, Kid, Old Man, Old Woman)
CNN [10]		
Proposed CNN		

표 5의 각 그래프의 좌측 y는 손실값, 우측 y는 정확도, x축은 회차(Epoch)이다. 또한, 파란색과, 초록색은 각각 훈련 셋과 검증 셋의 정확도, 노란색과 빨간색은 각각 훈련 셋과 검증 셋의 손실값를 나타낸다.

학습 초반에는 과적합 없이 두 개의 신경망 모두 정상적인 학습이 이루어지지만 학습이 지속될수록 [10]의 신경망은 훈련 셋과 검증 셋의 정확도와 손실값의 차이가 커지고 있으며, 과적합 현상이 크게 나타나고 있다. 과적합은 훈련 셋과 검증 셋의 정확도 차이가 클 경우로 판단할 수 있다.

제안하는 신경망이 모델 학습단계에서 더 높은 성능을 보였고 두 지표에서 정상적인 학습이 이루어졌다. 정확도는 [10]의 신경망에서 준수한 성능이 나타났지만, 과적합으로 인해 신뢰도가 떨어지고 5개 클래스 분류에서는 약 5%의 큰 성능차이가 나기 때문에 최종 성능평가에 앞서서 제안하는 신경망이 더 좋음을 판단할 수 있다.

최종 성능 평가에서 테스트 셋을 적용해 학습 단계에서의 성능에 대해 보다 명확하게 판단을 내리고 최종 결론을 도출한다.

Ⅳ. 성능 평가

성능평가는 테스트 셋의 분류 결과를 통해 여러 가지 성능지표를 도출하고 종합적으로 제안 모델의 성능을 평가한다.

테스트 셋에서의 지표별 성능과 혼돈매트릭스를 통해서 진행하며 모델의 성능(%)과 데이터셋에서 몇 개를 맞추었는지 살펴보고 최종 평가를 내린다.

우선, 테스트 셋에 대해 모델별로 정밀도(Precision), 재현율(Recall), F점수(F1-score)를 계산하여 분류 모델을 평가하였다[21] [22]. 3개 성능 지표에 대한 평가 결과는 표 6과 같이 나타났으며 동일한 클래스 수에서 모델별 성능은 제안하는 모델에서 더 높은 수준으로 나타났다.

표 6. 모델별 성능 평가 결과
Table 6. Performance evaluation results by model

Model	Class Number	Precision	Recall	F1-Score
CNN[2]	2	96%	96.5%	96%
	5	85%	83.6%	84.2%
Proposed CNN	2	97.5%	97.5%	97%
	5	91.6%	91.2%	91.6%

특히 F-점수는 정밀도와 재현율의 조화평균으로 계산된 성능 지표로 모델 평가지표로 많이 활용되며 여러 지표가 아닌 하나의 수로 평가를 내리는데 용이하다.

또한, 훈련 단계에서 계산된 검증 셋 정확도와 비교평가를 위해 테스트 셋의 정확도와 혼돈매트릭스를 계산하여 비교하였다.

표 7. 검증, 평가 데이터 셋 성능
Table 7. Validation, test data set accuracy

Model	Class Number	Validation set Accuracy	Test set Accuracy
CNN[2]	2	95.3%	96.2%
	5	85.5%	84.8%
Proposed CNN	2	97.1%	97.4%
	5	90%	91.4%

표 8. 5개 클래스에서의 모델별 혼돈 매트릭스
Table 8. Confusion matrix by model (only 5 class)

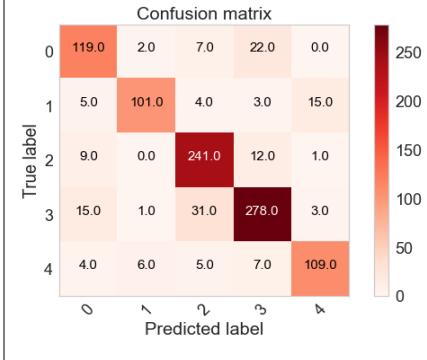
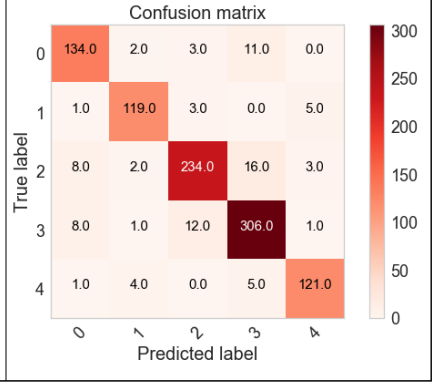
Model	Confusion matrix (5 Class)
CNN[10]	
Proposed CNN	

표 7에서 테스트 셋에 대한 정확도 성능 또한, 제안 모델에서 더 높게 계산되었다. 특히, 5개 클래스의 경우 테스트 셋에서 정확도 차이가 더 크게 나타났다. 이를 통해서 최종적으로 제안 모델이 더 정확하고 높은 성능을 낸다고 평가할 수 있다. 추가적으로 혼돈 매트릭스를 통해 모델이 어떤 클래스를 잘 분류하였는지 확인하였다.

표 8은 5개 클래스에 대한 혼돈 매트릭스이며 그래프의 y축 값은 0(여성), 1(남성), 2(어린이), 3(노인 여성), 4(노인 남성)는 각 클래스를 나타낸다.

2개 클래스의 경우 두 모델에서 높은 성능을 보였고, 클래스 수가 적기 때문에 분류 모델이 크게 혼동하는 경향을 보이지 않았다. 5개 클래스의 경우 여성과 남성의 경우 낮은 실패 수를 보였으나 여성, 어린이, 노인 여성에 대해서 상대적으로 실패하는 경우가 많이 나타났다. 이는 변성기시기에 명확한 차이를 보이는 남성과는 달리 여성은 목소리의 큰 변화를 보이지 않기 때문에 데이터셋의 특징 값이 비슷한 경향을 가지게 되어 변성기 이전의 어린이와 여성, 노인 여성을 혼동한다고 판단할 수 있다.

동일 환경에서의 학습 속도는 [10]의 신경망은 각 학습 회차(Epoch) 마다 2초 미만이었으며 제안하는 신경망에서는 약 3초대로 나타났다. 학습 속도 측면에서는 많은 회차를 학습할 때 개선의 여지가 있는 것으로 나타났다. 다만, 학습이 완료 된 모델을 통해 테스트를 하는 경우에는 1ms 미만으로 빠른 응답 속도를 보였으며 제품 내에서 모델을 지속적으로 학습하는 경우가 아니라면 학습된 모델을 그대로 활용하여도 문제가 없을 것으로 판단된다.

V. 결론 및 향후 과제

본 논문에서는 기존에 소리 분류에서 활용된 신경망[10]을 사람의 음성 분류에 적용하였을 때 성능을 확인하고 더 좋은 성능을 내는 모델을 제안하는 연구를 진행하였다. 기존 신경망[10]을 사람 음성 분류에 적용하였을 때, 클래스 수에 따라 큰 차이가 생기긴 했지만 비교적 분류를 잘 수행하였다. 하지만, 단순한 신경망 구성과 성능 측면에서 충분히 개

선될 여지가 있다고 판단하였으며 그에 따라 더 깊고 넓은 신경망 구성을 통해서 더 높은 성능의 모델을 제안하고 평가하였다.

사람 음성 분류의 경우 성인 남, 녀의 구분은 비교적 쉽고 단순한 신경망에서도 좋은 성능을 보이지만, 나이에 따라 여러 클래스를 분류 하는 것은 더 어려운 문제이다. 앞으로의 음성인식 분야에서는 성별 구분을 넘어서 연령대에 따른 맞춤형 서비스, 노래 추천 등으로 발전할 가능성이 높기 때문에 본 연구가 큰 의미를 가질 수 있다고 생각된다.

본 연구에서 5개 클래스를 분류할 경우 어린이, 여성, 노인 여성 부분에서 상대적으로 오 분류 하는 경향이 높게 나타났는데, 이 부분을 보완하기 위해서는 지금보다 더 많은 데이터 셋 구축과 전처리 방법 및 추가적인 신경망 개선을 통해 충분히 개선될 여지가 있다고 판단된다.

References

[1] H. S, Park, S. W. Kim, M. H. Jin, and C. D. Yoo, "Voice recognition technology trends based on machine learning", The Magazine of the IEIE, Vol. 14, No. 3, pp. 18-27, Mar. 2014.

[2] Deep Learning-based Speech Recognition Technology, http://www.itdaily.kr/news/articleView.html?idxno=76405. [accessed: Aug. 28, 2018]

[3] Filipp Akopyan et all., "TrueNorth: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip", IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Vol. 34, No. 10, pp. 1537-1557, Aug. 2015.

[4] A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio", arXiv:1609.03499v2, pp. 1-15, Sep. 2016.

[5] S. S. Jo and Y. G. Kim, "AI (Artificial Intelligence) Voice Assistant Evolving to Platform", IITP, pp. 1-25, Feb. 2017.

[6] mel-spectrogram, https://librosa.github.io/librosa/generated/librosa.feature.melspectrogram.html. [accessed: Aug. 28, 2018]

[7] M.D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks", arXiv: 1311.2901, pp. 819-833, Nov. 2013.

[8] L. H. Meng and J. S. Han, "The Impact of Relational Benefits on Positive Affect, Perceived Value, and Behavior Intention in Social Commerce : Focused on Chinese Tourist having the Hotel Service of Social Commerce environment", Journal of tourism and leisure research, Vol. 29, No. 10, pp. 69-88, Oct. 2017.

[9] J. H. Seo and Y. T. Kim, "Effects of Service Convenience on Customer Satisfaction and Reuse Intention by Korail Talk App Users among Korail Passengers", Journal of the Korean Society for Railway, Vol. 16, No. 5, pp. 410-417, Oct. 2013.

[10] H. Zhou, Y. Song, and H. Shu, "Using Deep Convolutional Neural Network to Classify Urban Sounds", TENCON 2017 - 2017 IEEE Region 10 Conference, pp. 3089-3092, Nov. 2017.

[11] Urban Sound DataSet, https://urbansounddataset.weebly.com. [accessed: Aug. 28, 2018]

[12] V. Nair and G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines", In Proc. 27th International Conference on Machine Learning, pp. 807-814, Jun. 2010.

[13] J. Schmidhuber, "Deep learning in neural networks: An overview", Neural Networks, Vol. 61, pp. 85-117, Oct. 2014.

[14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions", arXiv:1409.4842, pp. 1-12 Sep. 2014.

[15] M. S. Kim and T. W. Kang, "Proposal and Analysis of Various Link Architectures in Multilayer Neural Network", The Journal of Korean Institute of Information Technology, Vol.

16, No. 4, pp. 11-19, Apr. 2018.

[16] J. S. Choi, "A Speech Feature Enhancement Technique by Cepstral Noise Model for Noisy Speaker Identification", The Journal of Korean Institute of Information Technology, Vol. 16, No. 3, pp. 11-16, Mar. 2018.

[17] Keras, https://keras.io/. [accessed: Aug. 28, 2018]

[18] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical Evaluation of Rectified Activations in Convolution Network", arXiv:1505.00853, Nov. 2015.

[19] R. Caruana, S. Lawrence, and L. Giles, "Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping", in Proc. Neural Information Processing Systems Conference, pp. 402-408, Jan. 2000.

[20] Dobbin KK and Simon RM, "Optimally splitting cases for training and testing high dimensional classifiers", BMC Med Genomics, Apr. 2011.

[21] M. Sokolovaa and G. Lapalme, "A systematic analysis of performance measures for classification tasks", Information Processing & Management, Vol. 45, No. 4, pp. 427-437, Jul. 2009.

[22] A. Sattar and B. H. Kang, "Beyond Accuracy, F-score and ROC: a Family of Discriminant Measures for Performance Evaluation", AI 2006: AI 2006: Advances in Artificial Intelligence, Springer, pp. 1015-1021, Dec. 2006.

저자소개

박 대 서 (Dae-Seo Park)

2015년 8월 : 강원대학교 컴퓨터정보통신공학과 (공학사)
2018년 2월 : 강원대학교 컴퓨터정보통신공학과 (공학석사)
2018년 2월 ~ 현재 : 강원대학교 컴퓨터정보통신공학과 (연구과정생)
관심분야 : 데이터분석, 머신러닝, 딥러닝, 데이터공유 플랫폼

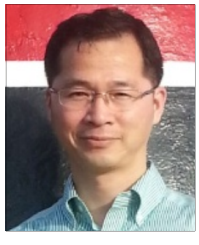
방 준 일 (Joon-Il Bang)

2018년 8월 : 강원대학교 컴퓨터정보통신공학과 (공학사)
2018년 9월 ~ 현재 : 강원대학교 컴퓨터정보통신공학과 (석사과정)
관심분야 : 데이터분석, 머신러닝, 딥러닝, 데이터공유 플랫폼

김 화 종 (Hwa-Jong Kim)

1982년 : 서울대학교 전자공학과 (공학사)
1984년 : KAIST 전기 및 전자과 (공학석사)
1988년 8월 : KAIST 전기 및 전자과(공학박사)
1988년 ~ 현재 : 강원대학교 컴퓨터정보통신공학과 교수
관심분야 : 데이터분석, 인공지능, 머신러닝, 딥러닝

고 영 준 (Young-Jun Ko)



1993년 2월 : 숭실대학교 전자계산학과(공학사)
1995년 2월 : 숭실대학교 전자계산학과(공학석사)
1994년 12월 : 대우정보통신 주임연구원
1999년 11월 : 모토로라 코리아 선임연구원
2005년 3월 ~ 현재 : 모다정보통신 IoT 개발단장
관심분야 : 데이터분석, 이동통신, IoT 플랫폼