JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 30 No. 10, pp. 53-60, October 2025
https://doi.org/10.9708/jksci.2025.30.10.053

A Lightweight CNN Model for Alcohol-intoxicated Detection Using Mel-Spectrogram Voice Features

Younguk Yun*

*Assistant Professor, Dept. of Software, Yonsei University, Wonju, Korea

[Abstract]

In this paper, we propose TinyAlcoCNN, a lightweight deep learning model designed to non-invasively detect alcohol consumption based on voice data. The proposed model adopts a 2D-CNN architecture that takes Mel-spectrograms as input and is trained on approximately 40,000 Korean voice samples. To support real-time applications, the dataset was preprocessed using the Whisper API for automatic segmentation. Experimental results demonstrate that TinyAlcoCNN achieves a training accuracy of 0.9982 and an inference accuracy of 1.000, while maintaining efficiency with approximately one million parameters and 13.9 million FLOPs. These results confirm both the effectiveness and computational efficiency of the model. This study highlights the feasibility of voice-based alcohol detection and suggests potential for broader applications, including personalized services, through multilingual expansion and integration with mobile systems.

▸ **Key words**: Alcohol Detection, Lightweight Deep Learning, CNN, Mel-Spectrogram, Edge Computing

[요 약]

본 연구는 음성 데이터를 기반으로 알코올 섭취 여부를 비침습적으로 판별할 수 있는 경량 딥러닝 모델인 TinyAlcoCNN을 제안한다. 제안 모델은 Mel-spectrogram을 입력으로 사용하는 2D-CNN 구조로, 약 30,000개의 한국어 음성 데이터를 활용해 학습되었다. Whisper API를 활용한 자동 분할을 통해 실시간 응용에 적합한 data set을 구성하였다. 실험 결과, 제안 모델은 0.9982의 학습 정확도와 1의 추론 정확도를 기록하며 높은 성능을 입증했고, 약 100만 개의 파라미터와 13.9M FLOPs로 경량성과 연산 효율도 확보하였다. 본 연구는 음성 기반 알코올 탐지의 가능성을 보여주며, 향후 다국어 확장 및 모바일 시스템 연계를 통해 실용성과 확장성을 더욱 강화할 수 있는 개인 맞춤형 서비스와 연구로 확장될 수 있다.

▸ **주제어**: 음주 상태 검출, 경량 딥러닝 모델, Mel-spectrogram, CNN, 엣지 컴퓨팅

• First Author: Younguk Yun, Corresponding Author: Younguk Yun
*Younguk Yun (yu_yun@yonsei.ac.kr), Dept. of Software, Yonsei University
• Received: 2025. 08. 21, Revised: 2025. 09. 23, Accepted: 2025. 09. 24.

Copyright © 2025 The Korea Society of Computer and Information
http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945
www.kci.go.kr

I. Introduction

최근 디지털 헬스케어 및 인공지능 기술의 발전은 단순한 질병 진단을 넘어서 비침습적이고 실시간 분석이 가능한 건강 모니터링 기술로 확장되고 있다. 특히, 음성 데이터는 비접촉 방식으로 사람의 감정 상태, 정신 질환, 신경학적 질병뿐만 아니라 행동 습관 및 중독 상태까지 파악할 수 있는 중요한 생체 신호로 간주되고 있다. 이러한 관점에서 음성 기반 데이터 분석 및 처리 기술은 향후 디지털 정신건강 관리 및 중독 예방 시스템의 핵심 구성 요소로 활용될 가능성이 높다. 특히 알코올 중독은 개인의 신체적 건강은 물론 사회적, 경제적 측면에서 심각한 영향을 미치는 만성 질환이며, 조기 발견과 개입이 치료 효과에서 매우 중요한 요소이다. 그러나 현재 널리 사용되는 진단 방법은 설문지 기반의 자기 보고, 생체검사(혈중알코올농도 등), 의사의 면담을 통해 진단하는 등의 객관성 부족 또는 침습적 방식에 기반하고 있어 일상 속에서의 지속적 감시는 현실적으로 어렵다. 따라서 생체 데이터를 통해 알코올 중독 여부를 판단할 수 있는 진단 도구가 필요한 시점이다.

한편, 최근에는 모바일 기기를 활용한 디지털 헬스케어 기술이 급속도로 발전하고 있다. 스마트폰, 웨어러블 디바이스, IoT 센서 등은 사용자의 건강 데이터를 실시간으로 수집하고 분석할 수 있으며, 언제 어디서나 건강 상태를 모니터링 가능한 인프라를 제공한다. 이들 모바일 기기는 카메라, 마이크, 가속도 센서, GPS 등 다양한 센서를 탑재하고 있어 음성, 영상, 움직임, 위치 정보 등 다양한 생체 및 행태 데이터를 수집할 수 있다, 기능 플랫폼으로 주목받고 있다. 특히 스마트폰은 거의 모든 인구가 일상적으로 사용하는 기기이기 때문에, 추가적인 의료 기기 없이도 건강 상태를 평가하거나 이상 징후를 감지할 수 있는 유용한 도구로 기능할 수 있다. 스마트폰을 이용한 알코올 중독 탐지와 예방 같은 정신, 행동 건강 문제에도 새로운 가능성을 열고 있다. 기존에는 병원이나 전문 센터에서만 가능했던 중독 평가 및 상담이, 모바일 기기를 통해 실시간으로 이루어질 수 있으며,

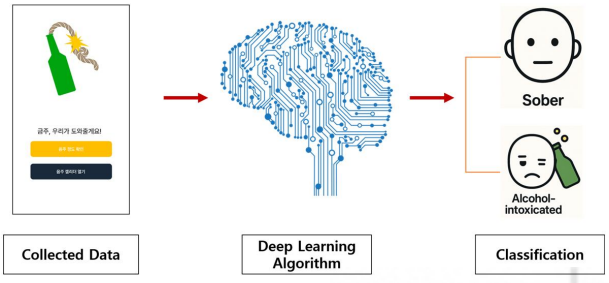

Fig. 1. System Description

그 결과를 바탕으로 사전 경고 및 개입도 가능해질 수 있다. 본 연구는 음성 데이터를 활용하여, 스마트폰과 같은 엣지 디바이스(edge-device)에서도 작동 가능한 경량 인공지능 모델을 제안한다. Fig. 1은 연구에서 제안된 모델이다. 추가적인 의료 기기 없이도 알코올 섭취 여부를 실시간으로 감지할 수 있으며, 사용자 맞춤형 디지털 헬스케어 시스템 구현에 활용될 수 있다.

II. Preliminaries

1. Related works

술을 섭취했을 때 반응으로 음성의 조음(articulation)에 많은 영향을 미치며, 특히 문장 전체와 특정 음소 단위 수준의 음향음운 특성에서 일관되고 명확한 속도의 변화가 나타난다. 말의 속도가 감소하고 전체적인 조음 조정 능력 저하 등이 특징이다. 이는 청취 실험과 디지털 신호처리 분석 모두에서 유의미하게 관찰된 현상이고, 정밀한 조음 조정 능력이 요구되는 음성 생산일수록 알코올의 영향이 더 크게 나타난다는 특징이 있다[1]. 숨을 불어 화학적 알코올 성분을 검출하는 생체검사 기반의 음주 측정기 방식 외 음성 신호 특징을 활용한 다양한 방식의 음주 측정 또는 판별 연구가 진행되고 있다.

기존 연구들은 주로 음성 특징 기반 탐지, 모바일/웨어러블 기반 실시간 분석, 시스템 융합 및 실생활 적용의 세 가지 방향의 연구가 진행되고 있다.

음성 데이터를 활용한 탐지 연구에서는 스펙트로그램, 주파수, jitter, shimmer 등 음향학적 특징을 추출하여 머신러닝 또는 딥러닝 기반 모델로 음주 여부를 예측하는 방식이 주를 이루고 있다.

Suffoletto et al.과 E. Bonelaet al.는 실험 참가자들이 특정 문장을 낭독한 음성 데이터를 수집한 후, 이를 스펙트로그램으로 변환하여 Support Vector Machine(SVM) 기반 분류 모델을 구축하였다[2-3]. 실험 결과, 혈중알코올농도(BrAC)가 0.08% 이상인 상태를 98%의 정확도로 분류할 수 있었으며, 민감도와 특이도는 각각 약 0.98, 0.97로 나타났다.

Bonela et al.는 Audio-based Deep Learning Algorithm to Identify Alcohol Inebriation(ADLAIA)라는 딥러닝 기반 모델을 제안하였다[4]. 이 모델은 독일어 음성 코퍼스에서 총 12,360개의 음성 클립을 활용하여 학습하였으며, 입력으로는 스펙트로그램이 사용되었다. 클래스의 탐지 능력을 고르게 평가해주는 지표인 balanced

www.kci.go.kr

accuracy(BAC)와 unweighted average recall(UAR)을 사용하고 있고 BAC가 0.05% 이상인 상태에서는 정확도 67.7%, BAC가 0.12% 이상에서는 unweighted average recall가 75.7%에 달했다. 본 연구는 딥러닝이 음주에 따른 음성 패턴의 변화를 인식할 수 있으며, 짧은 12초 음성 샘플(sample)만으로도 실시간 분석이 가능함을 보였다.

Terlapu와 Sadi는 모음 발성에 기반한 음성 샘플에서 frequency, jitter, shimmer, harmonic-to-noise ratio 등의 특징을 추출(feature extraction)하고, 이를 활용하여 Incremental Hidden Layer Neurons 기반 Backpropagation Artificial Neural Network (IHLN-ANN)를 제안하였다[5]. 실험에는 총 509개의 음성 샘플이 사용되었으며, 은닉 뉴런 개수를 2개에서 5개로 증가시키면서 성능을 비교하였다. 그 결과, 5개 은닉 뉴런 구성에서 99.4%의 정확도와 거의 1.0에 가까운 AUC를 기록하였으며, 정밀도(precision)는 99.66%, F1-score는 99.48%의 성능을 보였다.

또한 최근에는 구 목적과 적용 환경에 따라 경량화된 1D-CNN 기반 모델, 2D-CNN 기반 모델, 모든 특징을 동일하게 처리하지 않고 중요한 부분 또는 주파수에 대해서 큰 가중치를 부여해 학습을 수행하는 attention 모델이 결합된 확장 모델, 그리고 MLP 기반 모델로 구분할 수 있다.

우선, 1D-CNN 기반 모델은 원시 오디오 신호 혹은 1차원 feature sequence를 직접 처리하여 연산 효율성과 실시간성에서 강점을 보인다. 대표적으로 Al Badawi et al.이 제안한 AMMobileNet1D는 MobileNet 구조를 1차원 입력에 맞게 변형하여 임베디드 장치와 실시간 음성 분류에 적합하도록 설계되었다[6]. 또한, Venkataramani et al.의 연구는 Swish 활성 함수를 적용한 1D-CNN 구조로, 음성, 음악, 잡음을 효율적으로 분류할 수 있도록 최적화되었다[7].

한편, 2D-CNN 기반 모델은 Mel-spectrogram과 같은 2차원 입력을 활용하여 높은 범용성과 강력한 특징 추출 능력을 보인다. SpeechCNN_VGGSmall은 이미지 처리와 음성 신호 처리 관련 분야에서 강력한 성능을 보이는 대표 모델인 VGG 구조를 축소·경량화하여 음성 인식 및 분류에 적용되었으며[8], RACNN은 recurrent attention 알고리즘을 결합하여 세부 특징을 반복적으로 강조함으로써 분류 성능을 향상시켰다[9]. MOSLight는 음성 품질 평가(Mean Opinion Score, MOS)를 위한 경량 CNN 모델로 제안되었다[10]. 또한, AclNetLike는 원시 파형 입력을 직접 처리하는 end-to-end 방식의 2D-CNN 구조로, 87,026개에 파라미터(parameter)만 사용하여 모바일 환경에서도 사용 가능한 구조를 제안하였다[11].

최근에는 2D-CNN에 Attention 알고리즘을 결합한 모델도 연구되고 있다. LMFCA-Net은 multi-scale 특징 추출과 channel attention을 결합한 경량화 된 모델을 제안한다[12].

마지막으로, MLP 기반 모델은 특정 과제에 특화된 접근 방식을 제공한다. PIPMN은 피치 정보를 활용한 병렬 multi-branch 구조를 통해 음성 분리 및 잡음 억제 성능을 강화하였으며, CNN 기반 접근법과 차별화되는 대안을 제시하였다[13]. 이와 같이, 음성 처리 연구에서는 모델 경량화와 효율성 확보를 위한 다양한 시도가 이루어지고 있으며, 특히 모바일 및 엣지 디바이스 환경에서도 적용 가능한 모델 설계가 활발히 진행되고 있다.

기존 연구들은 음성 특징을 활용하여 알코올 상태를 인지하고 분류할 수 있음을 실험적으로 입증하였으며, 경량 모델 설계와 모바일 환경에서의 실시간 적용 가능성에 관한 연구도 활발히 이루어지고 있다. 그러나 선행 연구들은 대부분 영어 또는 타 언어를 대상으로 수행되어, 한국어 음성을 기반으로 한 분석 연구가 필요한 실정이다. 한국어는 교착어로서 조사와 어미 변화가 다양하고, 문장 종결이 모호하며 억양의 변화 폭이 작아 음성 분석 시 음주에 따른 변화가 미세하게 나타날 수 있다. 이에 본 연구는 한국어 음성의 구조적 특성을 반영하여, 전처리 과정을 수행하였다.

따라서 본 논문에서는 한국어 음성 데이터를 수집하고 알코올 섭취나 음주 여부를 판별하기 위한 딥러닝 연구를 수행하고 더 나아가 엣지 디바이스 환경에서도 구현 가능한 경량화 음성 처리 모델을 제안하고 분석한다. 이를 통해 개인화된 헬스케어 서비스로 활용될 수 있다.

III. The Proposed Scheme

1. Data Collection and Pre-processing

본 연구에서 활용된 데이터는 약 58시간의 음성 데이터이며 음성을 녹음하는 방식으로 데이터를 확보했다. 이중 절반은 정상 스크립트 녹음, 나머지는 음주 후 녹음을 진행했다. 참여자는 20대 남자 7명, 여자 5명이고 1시간 분량의 주어진 스크립트 읽는 방식으로 음성 데이터 수집했다. 녹음은 음주 이전 50분, 음주 후 50분 녹음은 임의로 지정한 두 날짜에 걸쳐 진행되었으며, 참여자 1인당 총 3일간 음주 전후의 음성을 수집하였다. 스크립트는 일상 대화와 뉴스 등 다양한 주제로 준비하고 50분이 초과하면 녹

음을 중지했다. 각 참여자는 다양한 스크립트 유형을 녹음했으며 스크립트 유형에 따라 감정, 속도의 변화를 유도하여 다양한 음성이 반영되도록 녹음을 진행했다. 술의 도수에 따라 취기가 올라오는 시간이 다르며 도수가 높을수록 음주 취기가 올라오는 시간이 더 빠르다는 연구[14-15]와 사람마다 만취까지의 주량 차이가 있을 수 있어 설문조사 후 가능한 최대의 30% 정도만 음주 후 30분 이후 녹음을 진행했다. 과적합이나 데이터 셋의 강건성을 높이기 위해 수음 환경은 실험자마다 동일 조건에서 진행되었다. Fig. 2와 같이 조용한 사무실 환경에서 수행하였으며 헤드셋을 활용하여 PC로 녹음이 이뤄지도록 환경 설정을 했다. 헤드셋의 제품명은 Shure WH20이며 오디오 인터페이스는 Tascam US-122 제품을 활용했다.

하나의 파일을 학습하는 것보다 녹음 파일을 여러 조각으로 나누는 것이 학습에 유리하다. 학습 샘플 수가 늘어나므로 데이터 부족 문제를 완화할 수 있고 짧게 잘라 학습하면 입력 일정한 길이가 모델이 일관된 학습 패턴을 얻을 수 있다. 또한, 실시간 처리 용이성 짧은 샘플 기반 학습은 추후 실시간 시스템 구현에도 적합하고 GPU 메모리 사용에도 유리하다. 반면 맥락(Context) 손실이라는 가장 큰 단점도 있는데 이를 보완하기 위해 7초 부근에서 문장이 끝나는 지점을 직접 확인하여 해당 시간 정보를 기록하고 이 정보를 기반으로 Whisper API를 활용해 데이터를 분리하는 전처리 과정을 거쳤다. 이 과정을 통해 약 7초 정도 시간으로 녹음 데이터를 분리해 정상(음주 전) 음성 파일과 음주 음성 파일 각 15,000개 파일, 약 30,000개 파일을 활용하여 딥러닝 분류학습을 수행하였다. Python에는 다양한 open source Automatic Speech Recognition(ASR) 엔진이 있는데 이 중 Whisper API를 활용해 파일을 분류하였다. Whisper 는 open-source API로 음성 인식 API로, 오디오 파일기능과 음성 파일을 분리할 때 사용할 때 사용할 수 있는 기능을 제공하고 있으며 타 API보다 접근성과 사용성이 좋아 주로 활용되는 tool이다.

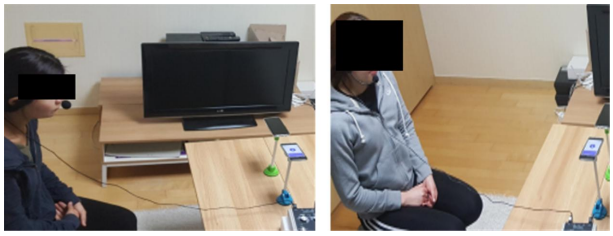

Fig. 2. Audio Recording Environment

2. Mel-spectrogram

음성 데이터는 다양한 잡음과 복합적인 신호가 혼합되어 있어, 원래의 의미 있는 정보를 효과적으로 추출하기 위해서는 신호 분리 과정이 필요하다. 이를 위해 본 연구에서는 Independent Component Analysis(ICA) 기법을 활용하여 서로 독립적인 원천 신호를 분리하고, 음성에서 중요한 특징만을 효율적으로 확보하고자 한다[16-17]. ICA는 혼합된 음향 신호 속에서 독립적인 통계적 특성이 있는 성분을 분리할 수 있기 때문에, 음성 인식이나 중독 탐지와 같은 응용 분야에서 노이즈 제거 및 특징 강조에 유용하게 활용된다. 추출된 원 신호는 이후 연산 효율성과 딥러닝 학습의 성능 향상을 위해 Mel scale 변환을 거친다[18-19]. 주파수 f(Hz)를 Mel 단위 m으로 변환하는 공식은 수식 (1)과 같다. Mel scale은 인간의 청각 인식 특성을 반영한 주파수 변환 방식으로, 특히 저주파 영역에서는 정밀하게, 고주파 영역에서는 상대적으로 단순화하여 표현한다. 이러한 변환은 음성의 지각적 특징을 보존하면서 불필요한 정보량을 줄이는 장점이 있으며, 이를 통해 생성된 Mel-spectrogram은 최근 음성 기반 인공지능 연구에서 표준 입력으로 자리 잡고 있다. Mel-spectrogram은 1차원 음성 신호를 2차원 신호로 변환한다. 딥러닝 학습 시 다양한 차원의 데이터를 제공하는 것이 학습에 유리하기에 이러한 형태로 음성 데이터를 가공하여 학습을 위한 data set(데이터 셋)으로 활용한다. Fig. 3.에서는 위에 그림은 정상 상태, 아래 그림은 취한 상태를 나타낸다. 시간에 따른 주파수 데이터로 x축은 시간, y축은 주파수를 나타낸다.

$$m = 2595 \cdot log_{10}(1 + \frac{f}{700}) \qquad (1)$$

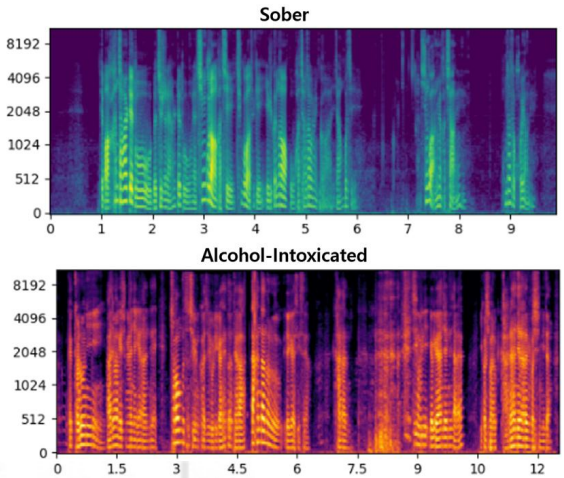

Fig. 3. Example of Mel-spectrogram

www.kci.go.kr

3. Proposed Deep-learning Algorithm

딥러닝 모델학습을 위해 원본 음성 파일은 전처리 과정을 거쳐 Mel-spectrogram 형태로 변환한다. 각 데이터는 128 × 128의 입력 형태로 정규화한다. 이 과정에서 입력 길이가 부족한 경우에는 0의 값을 넣는 zero-padding을 적용하여 일정한 크기를 유지하고, 길이가 초과되는 경우에는 cropping을 수행하여 딥러닝 모델 학습을 위한 일관성있는 데이터셋으로 변환한다. 최종적으로 각 Mel-spectrogram은 (1, 128, 128) 차원의 텐서 형태로 변환되며, 해당 데이터와 대응되는 클래스 레이블(label)이 함께 반환된다. 이렇게 구축된 데이터셋은 제안하는 모델의 입력으로 사용되며, 음성 데이터의 time-frequency 특징을 CNN 구조에서 효과적으로 학습할 수 있도록 한다.

본 연구에서는 제안하는 모델의 구조는 Convolutional Neural Network(CNN) 기반의 경량 구조를 제안한다. 제안하는 TinyAlcoCNN 모델은 연산량을 최소화하면서도 음성 데이터의 주요 특징을 효과적으로 추출할 수 있도록 구현했다. 해당 모델은 크게 convolutional(합성곱) Block과 fully-connected block으로 구성된다.

먼저, 합성곱 블록은 두 개의 합성곱 층과 정규화 및 활성화, pooling 연산으로 이루어진다. 첫 번째 합성곱 계층은 입력된 1채널 음성 spectrogram을 8개의 필터로 변환하며, batch normalization과 ReLU 활성화 함수를 통해 학습의 안정성과 비선형성을 확보한다. 이어서 max-pooling을 적용하여 공간 차원을 절반으로 축소함으로써 연산 효율성을 높인다. 두 번째 합성곱 계층은 16개의 필터를 활용하여 더욱 복잡한 특징을 추출하며, 동일한 fully-connected 블록은 Flatten 연산을 통해 추출된 특징 맵을 1차원 벡터로 변환한 후, 64차원 hidden layer를 거쳐 최종적으로 2차원 output layer(출력층)를 구성한다. Fig. 4는 제안하는 모델이다. 출력층은 음성 데이터가 정상 음성과 음주 음성, 두 가지 클래스 중 어느 범주에 속하는지를 구분하는 이진 클래스로 구현되었다.

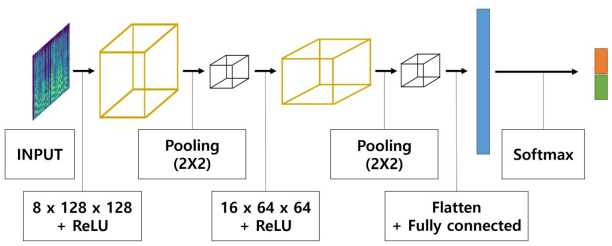

Fig. 4. Proposed TinyAlcoCNN Algorithm

제안하는 TinyAlcoCNN은 소규모 파라미터수와 낮은 연산 복잡도를 특징으로 한다. 따라서 일반적인 대규모 딥러닝 모델에 비해 메모리 사용량과 계산 시간이 크게 줄어들며, 이는 모바일 및 엣지 디바이스와 같은 제한된 자원 환경에서의 실시간 음성 인식 및 분류 적용 가능성을 높인다. 또한, 구조적 단순성을 유지하면서도 합성곱 연산과 정규화, 활성화의 조합을 통해 기본적인 음성 특징 추출 및 분류 성능을 확보할 수 있는 구조로 설계되었다.

4. Experimental Results and Evaluation

Table 1은 본 논문에서 비교한 모델들의 주요 성능 지표들을 요약한 것이다. 비교 항목에는 표의 순서대로 model(모델)은 II. 1. 절에 소개된 경량 딥러닝 학습모델이고 family는 딥러닝 모델의 계열을 의미한다.

표에서 보는 것처럼 제안하는 모델 외 비슷한 계열을 묶어 순서대로 나열했다. Training Loss(학습 손실), Training Accuracy(학습 정확도), Params(파라미터 수), FLOPs(계산 복잡도), train time(학습 시간), inference accuracy(추론 정확도), inference time(추론 시간)이 포함된다. table 항목 괄호의 의미는 M은 mega, S는 second이다. 9번 모델인 TinyAlcoCNN은 본 논문에서 제안하는 구조이다. 학습 시 learning late는 1e-3, epoch(반복)는 10회로 진행 후 결과를 분석했다. 실험 환경은 제안하는 모델은 연산 복잡도가 낮고 엣지 디바이스 실험 환경에서 환경에 서도 충분히 적용 가능함을 보이기 위해 사무용 PC 환경에서 학습을 진행했다. CPU는 I사의 i7-10700(2.90GHz), GPU는 N사의 GTX 1660 Super(6GB RAM)에서 진행되었다. 추론 정확도와 시간은 모든 데이터에 대해 2,000개의 sample(샘플) 데이터를 임의로 추출했다. 여기서 1,000개는 정상상태(sober), 1,000개는 음주 상태(alcohol-intoxicated) 상태이다. Fig. 5는 학습 시간과 학습 손실에 대한 그림이다. 실험 결과 제안하는 TinyAlcoCNN은 학습 손실 0.0041, 학습 정확도 0.9987를 기록하며, 빠르고 안정적인 수렴 특성을 보였다. 이는 LMFCA_Net, MOSLight, RACNN과 유사한 수준으로, 학습 최적화가 충분히 이루어졌음을 의미한다. 제안 모델은 1,050,066개의 파라미터와 13.9M FLOPs로, 추론 정확가 0.998으로, 비교 대상 모델 중 가장 경량이며 계산 효율이 높다. RACNN은 유사한 정확도를 보이지만 FLOPs가 32.76M으로 약 2.4배 높다. PIPMN은 FLOPs는 비슷하지만, 파라미터 수가 약 17배 많다.

www.kci.go.kr

Table 1. Experimental Results Table

No.	model	Family	Training Loss	Training Acc.	Params	FLOPs (M)	Training Time (s)	Infer. Acc.	Infer. Time (s)
1	AMMobileNet1D	1D-CNN	0.1520	0.9436	26,722	3.94	624.57	0.934	1.59
2	SwishNet	1D-CNN	0.6192	0.6607	13,378	2.17	624.4	0.577	1.64
3	LMFCA Net	2D+Attn	0.0115	0.9970	4,244,834	20.25	663.75	0.996	1.2
4	AclNet	2D-CNN	0.2384	0.8985	87,026	0.58	820.92	0.910	1.18
5	MOSLight	2D-CNN	0.0099	0.9966	2,102,146	46.66	678.95	0.980	1.09
6	RACNN	2D-CNN	0.0055	0.9982	1,050,602	32.76	682.17	0.995	1.07
7	SpeechCNN VGGSmall	2D-CNN	0.0019	0.9994	139,746	915.40	1196.67	1.000	3.23
8	PIPMN	MLP	0.6932	0.4985	17,336,002	34.67	630.32	0.449	0.88
9	*TinyAlcoCNN	2D-CNN	0.0041	0.9987	1,050,066	13.90	676.73	0.998	1.07

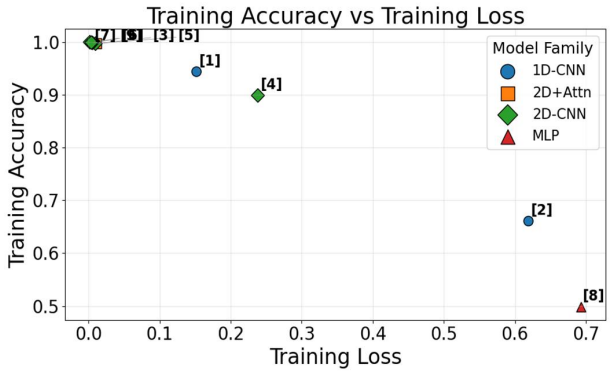

Fig. 5. Training Accuracy and Loss

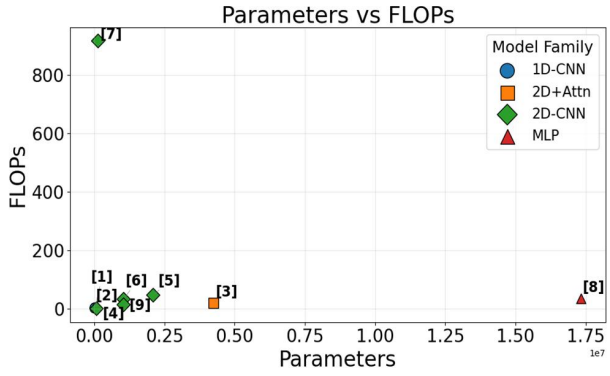

Fig. 6. Number of Parameters and Flops

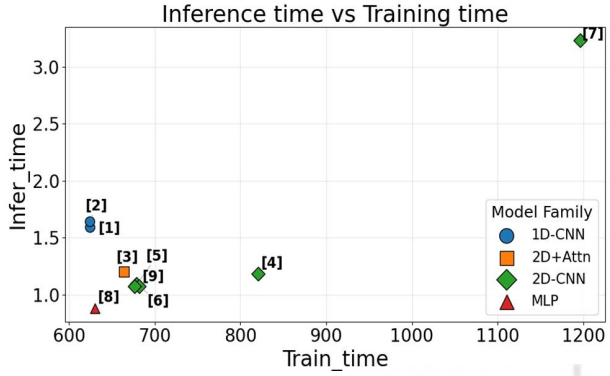

Fig. 7. Inference time and Training time

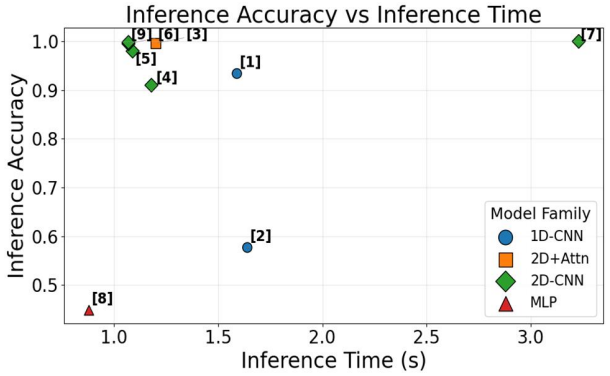

Fig. 8. Inference time and Inference accuracy

SpeechCNN VGGSmall은 VGG 계열 모델 중에서 경량에 속하지만 FLOPs가 915.40M으로 제안 모델보다 66배 이상 크고, 추론 시간도 3초 이상으로 보였다. Fig. 6는 각 모델의 파라미터 수와 FLOPs의 관계를 나타낸다. 파라미터 수와 FLOPs는 학습이나 추론 속도에 영향을 미치는 지표이지만 실제 학습 시간 시 FLOPs에 대한 수치가 더 많은 영향을 미친다고 볼 수 있으며 두 지표가 항상 선형적인 관계를 보이는 것은 아니다. 이는 convolution, attention 등 모델 구조에 따라 동일한 파라미터 수라도 연산량이 달라질 수 있기 때문이다.

Fig. 7은 추론 시간과 학습 시간의 관계, Fig. 8은 추론 정확도와 추론 시간의 관계를 나타낸다. 학습 및 추론 시간은 제안 모델의 경우 학습 시간은 676.73초, 추론 시간은 1.07초로 매우 효율적이다. 학습 시간은 다른 모델보다 압도적으로 우수하다고 볼 수 없으나 학습 정확도가 높은 LMFCA Net, MOSLight, RACNN 모델들과 유사하며, 추론 시간은 추론 시간은 약 1.07초로 매우 짧아 실시간 응용이 가능함을 보여준다. 예를 들어, SpeechCNN VGG Small의 추론 시간은 3.22초로 TinyAlcoCNN 대비 3배 이상 느리다.

www.kci.go.kr

본 논문에서 제안 모델인 TinyAlcoCNN의 실험 결과 낮은 손실과 높은 학습 정확도를 보여 탁월한 학습 수렴의 특성을 파악했다. 이진 분류이지만 최고 수준의 추론 정확도를 보였으며 실시간/경량화 시스템에 최적화 모델로 활용될 수 있음을 확인했다. 본 연구는 개별 사용자 단위에서의 정밀한 음성 변화 탐지에 초점을 맞춘 연구로도 활용될 수 있다.

IV. Conclusions

본 연구에서는 음성 데이터를 기반으로 알코올 섭취 여부를 판별할 수 있는 경량 딥러닝 모델인 TinyAlcoCNN을 제안하고, 이를 모바일 및 엣지 디바이스 환경에서도 실시간으로 적용 가능함을 실험적으로 입증하였다. 제안된 모델은 학습 정확도 0.9987와 추론 정확도 0.998로 100%에 가까운 정확도를 달성하며 기존 경량 모델들과 비교해 높은 정확도와 안정적인 학습 수렴 특성을 보였다. 또한, 1백만 개 수준의 파라미터 수와 13.9M FLOPs라는 낮은 연산 복잡도를 통해 연산 자원 소비를 최소화하면서도, 실시간 처리에 필요한 속도와 정확도를 모두 만족시키는 결과를 도출하였다. 이러한 경량성과 효율성은 특히 자원이 제한된 모바일 디바이스나 임베디드 환경에서 음성 분석 시스템을 실용적으로 구현할 수 있는 가능성을 보였다.

본 연구를 통해 별도의 생화학 검사나 의료 장비 없이, 스마트폰이나 마이크가 내장된 웨어러블 기기 등 일상적인 디지털 기기를 활용해 비침습적이고 자연스러운 방식으로 알코올 중독 상태를 감지할 수 있는 가능성을 확인했다. 이는 기존의 설문지 기반 자기 보고, 혈중알코올농도 측정기, 의료진의 면담 등과 같은 침습적이고 일회성의 진단 방식이 가진 한계를 넘어서는 접근으로, 디지털 정신건강 관리와 행동 중독 예방 분야에서 실질적인 대안이 될 수 있다. 특히 음성 인식, 감정 분석, 중독 탐지 등 다양한 디지털 헬스케어 분야로의 확장이 가능하다는 점에서 높은 응용 가치를 가진다. 본 연구는 개별 사용자 단위에서의 정밀한 음성 변화 탐지에 초점을 맞춘 연구에 더 적합할 수 있다.

향후 연구에서는 한국어 외 다양한 언어와 방언을 포함한 데이터 셋 확장, 성별, 연령, 발화 습관 등 다양한 사용자 특성을 고려한 모델 정교화, 그리고 음주 여부를 이진 분류하는 단계를 넘어 경도, 중등도, 고도 취기와 같은 다단계 분류 체계 도입이 필요하다. 또한, 사용자와의 상호작용을 기반으로 실시간 경고 및 피드백 기능이 포함된 모바일 애플리케이션과의 연동을 통해 실제 생활 환경에서 적용 가능한 통합 중독 관리 시스템으로 발전시킬 수 있을 것이다. 또한, 사용자별 음성 베이스라인을 학습하는 방식으로 확장하여 개인 맞춤형 서비스로 활용될 수 있다.

REFERENCES

[1] D. B. Pisoni and C. S. Martin, "Effects of Alcohol on the Acoustic-Phonetic Properties of Speech," Journal of Studies on Alcohol, Vol. 13, No. 4, 577-587, 1989. DOI: 10.1111/j.1530-0277.1989.tb00381.x

[2] C. Suffoletto, A. Akhtar, E. Bonela, A. Dey, and K. A. Vemana, "Detection of Alcohol Intoxication Using Voice Features: A Controlled Laboratory Study," Journal of Studies on Alcohol and Drugs, Vol. 84, No. 2, pp. 241-247, Mar. 2023. DOI: 10.15288/jsad.22-00375.

[3] C. Suffoletto et al., "Smartphones and Smart Speakers Can Detect Alcohol Intoxication From Voice," MedicalXpress, https://medicalxpress.com/news/2023-11-smartphones-smart-speakers-alcohol-intoxication.html

[4] E. Bonela, C. Suffoletto, and A. Dey, "Audio-Based Deep Learning Algorithm to Identify Alcohol Inebriation," Computers in Human Behavior, Vol. 129, 107123, Dec. 2022. DOI: 10.1016/j.alcohol.2022.12.002.

[5] S. Terlapu and K. Sadi, "Real-Time Speech-Based Intoxication Detection System: Vowel Biomarker Analysis with Artificial Neural Networks," International Journal of Advanced Computer Science, vol. 35, no. 2, pp. 55-64, 2024. DOI: 10.12785/ijcds/1501116

[6] Nunes, J. A. C., Macêdo, D., and Zanchettin, C. "Am-mobilenet1d: A portable model for speaker recognition," Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), IEEE, 2020. DOI: 10.1109/IJCNN48605.2020.9207519

[7] Hussain, M. S., and Haque, M. A. "Swishnet: A fast convolutional neural network for speech, music and noise classification and segmentation," arXiv preprint arXiv:1812.00149, 2018. DOI: 10.48550/arXiv.1812.00149

[8] Simonyan, K., and Zisserman, A. "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014. DOI: 10.48550/arXiv.1409.1556

[9] Fu, J., Zheng, H., and Mei, T. "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4438-4446, 2017. DOI: 10.1109/CVPR.2017.476

[10] Li, Z., and Li, W. "MOSLight: A Lightweight Data-Efficient System for Non-Intrusive Speech Quality Assessment,"

www.kci.go.kr

Proceedings of the Interspeech 2023, pp. 5386-5390, 2023. DOI: 10.21437/Interspeech.2023-263

[11] Huang, J. J., and Leanos, J. J. A. "Aclnet: efficient end-to-end audio classification cnn," arXiv preprint arXiv:1811.06669, 2018. DOI: doi.org/10.48550/arXiv.1811.06669

[12] Zhang, Y., et al. "LMFCA-Net: A Lightweight Model for Multi-Channel Speech Enhancement with Efficient Narrow-Band and Cross-Band Attention," Proceedings of the ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2025. DOI: 10.1109/ICASSP49660.2025.10889867

[13] Chen, Y., Zhu, Y., Yan, Z., Ren, Z., Huang, Y., Shen, J., and Chen, L. "Effective audio classification network based on paired inverse pyramid structure and dense MLP Block," Proceedings of the International Conference on Intelligent Computing, pp. 70-84, Singapore: Springer Nature Singapore, July 2023. DOI: 10.1007/978-981-99-4742-3_6

[14] M. C. Mitchell Jr et al., "Absorption and peak blood alcohol concentration after drinking beer, wine, or spirits," Alcoholism: Clinical and Experimental Research, Vol. 38, No. 5, 1200−1204, 2014. DOI: 10.1111/acer.12355

[15] A. Paton, "Alcohol in the body," BMJ Study, 2005. DOI: 10.1136/bmj.330.7482.85

[16] J.-H. Lee, H.-Y. Jung, T.-W. Lee, and S.-Y. Lee, "Speech Feature Extraction Using Independent Component Analysis," Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Vol.3, pp. 1631-1634, Istanbul, Turkey, June 2000. DOI: 10.1109/ICASSP.2000.862023

[17] D. Kolossa, ""Independent component analysis and time-frequency masking for speech recognition in multitalker conditions," EURASIP Journal on Audio, Speech, and Music Processing EURASIP Journal on Advances in Signal Processing, Vol. 2010, No. 651420, pp. 1-14, Jan. 2010. DOI: 10.1155/2010/651420.

[18] Fischer, T., Schneider, J., & Stork, W. "Classification of breath and snore sounds using audio data recorded with smartphones in the home environment," IEEE International Conference on Acoustics, Speech and Signal Processing, May. 2016. DOI: 10.1109/ICASSP.2016.7471670

[19] Lane, N. D. Georgiev, P. & Qendro, L., "Deepear: robust smartphone audio sensing in unconstrained acoustic environments using deep learning," In Proceedings of the ACM international joint conference on pervasive and ubiquitous computing, pp. 283-294, September 2015. DOI: 10.1145/2750858.2804262

Authors

Younguk Yun received the B.S. degree in 2014 and the integrated M.S./Ph.D. degree in 2020, both in Electronic Engineering from Kwangwoon University, Korea. Since 2021, he has been with the Department of Software

He is currently an Assistant Professor with the Department of Software, Yonsei University, South Korea. His research interests include artificial intelligence and its applications to indoor positioning, healthcare, and autonomous driving, with a focus on signal processing using vision, radar, IMU, and IoT sensors.

www.kci.go.kr