# ResNet 모델을 이용한 일상생활 소리 예측 및 알림 애플리케이션

박유진*, 정은이*, 신지혜*, 박태정*, 양회석**
*덕성여자대학교 IT 미디어공학과
**나이스컨설팅
vivaeugene09@duksung.ac.kr, chungee375@duksung.ac.kr, jjihae1998@duksung.ac.kr, tjpark@duksung.ac.kr, impperfume@naver.com

# ResNet Model Based Real Life Sound Event Prediction and Notification Application

Yu-Jin Park*, Eun-Ee Chung*, Ji-Hye Shin*, Tae-jung Park*, Yang-Hoi Seok**
*Dept. of IT Media Engineering, Duksung Women's University
**Nice Consulting

**요 약**

본 논문에서는 청각 장애인이 가정에서 듣지 못해 발생하는 낭비와 위험을 미리 예방하기 위하여 가정에서 현재 발생하고 있는 소리를 알려주는 시스템을 구현하였다. 무지향성 마이크로 일상 소리 감지 후 음향 데이터에서 Mel-Spectogram 특징 벡터를 추출하여 Convolutional Neural Network (CNN) 모델의 Resnet 알고리즘을 진행한다. 서버에서 소리에 대한 분석을 진행한 후 그 결과를 안드로이드에서 실시간으로 5 초마다 확인하여 사용자에게 알림 서비스를 제공한다. 이를 통해 낭비를 줄이고 위험에 대처할 수 있게 한다. 청각 장애인의 소리에 대한 접근성을 다양한 측면으로 고려해야 한다는 사회적 인식을 확산시키고자 한다.

## 1. 서론

청각 장애인은 가정에서 듣지 못하여 발생하는 상황을 마주한다. 들을 수 없기 때문에 수돗물이 계속 나와도 알지 못하며 드라이기에 직접 손을 대보고 켜져 있는지 확인해야 한다. 이렇게 몰라서 생긴 낭비가 발생한다.[1]

또한, 모든 판단을 시각에 의존해야 한다. 청인보다 수용할 수 있는 정보량이 제한적이고 정보의 접근성이 낮아 위험 상황에서 시간상 너무 늦게 정보를 획득하거나 정보를 전혀 얻지 못하는 경우가 있다. 그로 인해 위와 같은 낭비가 발생하기도 하고, 위험 상황에 더 쉽게 노출된다.[2][3]

본 논문에서는 청각 장애인을 위하여 가정에서 발생하는 소리를 인식하고 추론하여 현재 소리 발생 상황을 알려주는 애플리케이션 Soundee 를 개발하고자 한다. 무지향성 마이크로 입력된 음향신호에서 Mel-Spectrogram 특징 벡터를 추출하여 패턴으로 이미지를 분류하는 Convolutional Neural Network (CNN) 모델의 Resnet 50 구조를 사용하여 발생한 음향 소리를 예측한다.

## 2. 관련 기술 현황과 차별성

### 2.1 소리우산

유사 애플리케이션 "소리우산"은 스마트 밴드 착용이 필요하며 데시벨마다 소리 알림을 다르게 전달한다. 데시벨 측정으로 데시벨 단계마다 알림을 주는 애플리케이션과는 달리 본 논문에서는 소리 신호에서 추출한 특징 벡터를 바탕으로 CNN 모델의 Resnet 알고리즘을 사용하여 소리를 예측하였다. 예측한 소리 종류를 사용자에게 알려줌으로써 같은 데시벨이라도 다른 종류의 소리 정보 분류가 가능하여 더 디테일한 소리 정보 제공이 가능하게 하였다.

### 2.2 IoT 상품

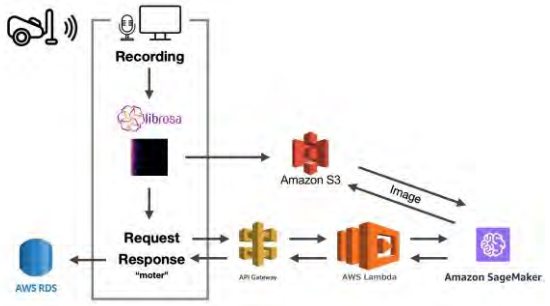
시중에 나와 있는 IoT 상품 중 가전제품 전력 낭비를 줄이기 위해 전원을 제어할 수 있는 것이 있다. 그러나 해당 상품과 연결되는 가전제품만 가능하며 기기를 설치해야 하고, 달마다 요금이 부과된다. 요금으로 부담이 있을 사용자를 위해 본 논문에서는 유지 비용이 들지 않는 마이크 장치와 애플리케이션 설치만으로 가정의 상황을 파악할 수 있게 하

려 한다.

## 3. 설계 및 구현

### 3.1 프로젝트 구조도

그림 1 와 그림 2 는 각각 본 논문의 애플리케이션과 이를 위한 녹음 프로그램의 절차를 나타낸 프로젝트 구조도이다.




(그림 1) Soundee Recorder 와 서버 서비스 구조도

Soundee Recorder 는 로그인을 통해 사용자 정보를 얻고 5 초 단위로 녹음이 실행된다. 녹음된 오디오 데이터는 Librosa Library 를 통해 이미지로 전처리 되어 사용자 정보를 담아 AWS S3 에 업로드된다. 이후 AWS API Gateway 에 요청을 보내어 Lambda 를 통해 Sagemaker 엔드포인트에 요청을 보낸다. Sagemaker 는 S3 에서 요청을 보낸 사용자 정보에 맞는 이미지를 추론하여 결괏값을 응답한다. 응답한 결괏값은 Recorder 가 받아 사용자 정보, 시간과 함께 즉시 AWS RDS 에 저장된다.


(그림 2) Soundee Application 구조도

본 논문의 녹음 프로그램과 애플리케이션을 동시에 실행시킬 때, 애플리케이션에서는 5 초 마다 서버에 요청을 보낸다. 서버는 요청받은 시간부터 5 초 전까지의 데이터를 조회하여 데이터가 있는 경우, 현재 소리가 추론하는 class 에 대하여 유의미한 소리임을 판단한다. 데이터가 없는 경우, 현재 소리가 추론하는 class 에 대하여 무의미한 소리임을 나타낸다. 이에 애플리케이션에서는 실시간으로 나는 소리 class 를 파악할 수 있다. 클라이언트는 해당 소리를 받아 class 별로 팝업 알림을 띄워 사용자에게 소리를 알려준다.

녹음 프로그램과 애플리케이션 각각 5 초 단위로 데이터 삽입과 조회가 이루어지기 때문에 실시간으로 발생하는 소리 정보를 얻어낼 수 있다.

### 3.2 DB 설계

그림 3 은 본 논문의 애플리케이션 구현을 위해 사용한 데이터베이스의 관계를 나타낸 관계도이다.


(그림 3) 데이터베이스 관계도

Sound info, User info, Sound class 세 개의 데이터베이스를 가지고 Sound info 와 User info 는 서로 관계를 맺고 있다.

그림 4 는 Database 들의 정의를 나타내고 있다.

| 테이블명 | 필드명 | DATA TYPE | 데이터길이 | KEY | NULL고려유무 | 자동증분 | 디폴트값 |
|---|---|---|---|---|---|---|---|
| sound | soundIdx | int | int(11) | PRI | NO | auto_increment | NULL |
| sound | class | varchar | varchar(45) | | NO | | NULL |
| sound | eventdate | datetime | datetime | | NO | DEFAULT_GENERATED | CURRENT_TIMESTAMP |
| sound | sound_userIdx | int | int(11) | MUL | NO | | NULL |
| sound_class | classIdx | int | int(11) | PRI | NO | auto_increment | NULL |
| sound_class | class | varchar | varchar(45) | | NO | | NULL |
| user | userIdx | int | int(11) | PRI | NO | auto_increment | NULL |
| user | email | varchar | varchar(45) | UNI | NO | | NULL |
| user | password | varchar | varchar(200) | | NO | | NULL |
| user | salt | varchar | varchar(200) | | NO | | NULL |
| user | name | varchar | varchar(10) | | NO | | NULL |

(그림 4) 테이블 명세서

Sound 는 녹음 프로그램에서 얻은 현재 소리에 대한 인덱스, 소리 종류, 발생 시각, 소리에 대한 사용자 정보에 대한 데이터베이스 테이블 구성이다.

Sound_Class 는 소리 종류에 대한 인덱스, 소리 종류 명에 대한 데이터베이스 테이블 구성이다.

User 는 사용자에 대한 인덱스, 이메일, 비밀번호, 솔트 값, 이름에 대한 데이터베이스 테이블 구성이다.

### 3.3 음향 예측 알고리즘

Convolutional Neural Network (CNN) 모델을 기반으로 한 딥러닝 소리 예측을 하기 위해선 아래와 같이 크게 3 가지의 단계가 있다.

#### 3.3.1 음향 데이터 수집

딥러닝으로 일상생활 소리를 예측하기 위함이므로 소리에 대한 학습이 필요하였다. 가장 우선시 되는 준비는 데이터 셋이다. 종류는 물 흐르는 소리, 모터 소리, 아기 울음소리, 사이렌 소리 4 가지로 구성하였다. 다양한 출처를 통해 데이터를 수집하였다. DCASE 2018 task5 의 공개 데이터 일부를 사용하였다. [4]

Kaggle 데이터 중 Environmental Sound Classification 50 에서 데이터를 수집하였다. [5]

무료 저작권 사이트 Freesound 에서 데이터를 수집하였다. [6]

표 1. 음향 이벤트 클래스

| Event class label | Number of instances |
|---|---|
| Baby crying | 1866 |
| Motor | 2545 |
| Siren | 1099 |
| Water | 2990 |

### 3.3.2 데이터 정규화와 음향 데이터의 특징 추출

수집한 데이터 셋의 크기는 DCASE 10 초, Kaggle 5 초, Freesound 는 약 6 초에서 길게는 10 분이 넘어가는 등 일정하지 않았기 때문에 이를 정규화하는 작업이 필요하였다. 어느 정도 정제 되어있는 Kaggle 데이터를 기준으로 하여 음향 데이터 길이를 5 초로 정하여 5 초 이상이면 잘라주는 방식으로 정규화를 진행하였다.

음향 데이터 예측을 진행하기 위해서 음향의 특징을 추출해야 했고 음악과 소리 분석 라이브러리 librosa 를 사용하였다. 특징 추출에는 mfcc, mel-spectrogram 2 가지 방법이 있다.

mel-spectrogram 은 주파수 단위를 멜 단위(사람이 듣는 소리에 대한 단위)로 바꾼 스펙트럼이고, mfcc 는 mel scale spectrum 을 40 개의 주파수 구역으로 묶은 뒤에 이를 다시 푸리에 변환하여 얻은 계수이다. 음성 인식에서 가장 널리 사용되고 있는 알고리즘이며 벡터화되어 있다는 것이 강점이다.

참고한 논문[7]에 따라 1 단계로 mel-spectrogram, 2 단계 filter banks 로 하여 특징 벡터를 추출하였다.

본 논문에서는 녹음한 16kHz 의 입력 오디오 데이터를 40ms 윈도우 단위로 0.01 초 마다 frequency 를 뽑은 melspectogram 을 추출하였으며 이에 대하여 mel 값을 얻어내기 위한 filter bank 를 적용하여 2 차원 입력 이미지 데이터를 구성하였다.

### 3.3.3 CNN Resnet 기반 음향 이벤트 인식 알고리즘 학습

기존에는 논문[7]의 하이퍼 파라메터를 참고하여 3 개의 convolutional layer 와 2 개의 fully-connected layer 로 구성하였다. tensorflow 로 CNN 알고리즘을 구성하여 진행하였으나 학습 정확도가 저조하였다. network 의 깊이가 얕아서 발생한 문제라고 생각되어 CNN 알고리즘에서 각광받는 ‘Resnet’ 구조를 사용하기로 하였다.

본 논문에서는 AWS sagemaker 에서 학습을 진행하였다. AWS sagemaker 에서 딥러닝을 하기 위해서는 ‘.pkl’ 파일이나 ‘.rec' 파일이 권장되었다. label 이 들어간 csv 파일을 만들고 이미지를 (256, 256) 크기로 일괄적으로 변경한 후 데이터 pickling 과정을 거쳐 ‘.pkl’ 파일을 생성하였다. ‘Apache MXNet’ 에서 제공하는 방법으로 ‘.rec' 파일을 생성하였다. 두 파일을 비교하여 '.rec' 파일이 더 가볍게 만들어진 것을 확인하고 최종적으로 딥러닝에 '.rec' 파일을 사용하기로 하였다.

Resnet layer 50 개, image shape (3, 256, 256), class 4 개, epoch 20 으로 설정하였고 학습률(learning rate)은 0.00001 에서 1.0, 미니 배치(mini batch) 크기는 16 에서 32, 최적화 알고리즘은 SGD, Adam, RMSprop, nag 네 가지에서 자동으로 사용되게 구성하였다.

early stopping 기능을 추가하여 일정 횟수 동안 손실(loss) 값의 개선이 이루어지지 않으면 학습을 중단하도록 하였다.

9 번째 epoch 에서 early stopping 되었고, 정확도 약 0.94693 으로 학습이 마무리되었다.

### 3.4 애플리케이션 구현

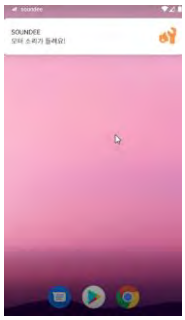
그림 5 은 애플리케이션 화면 흐름도이다. 핵심적인 기능은 듣는 소리를 나타내주는 홈 뷰와 일일/주간/월간으로 어떤 소리가 났는지 그래프로 표시해주는 통계 뷰이다.


(그림 5) 애플리케이션 화면 흐름도

사용자는 로그인하면, 백그라운드에서 5 초마다 현재 소리에 대한 데이터를 요청한다. 특정 Class 에 대한 소리가 있다는 응답을 받으면 그림 6 과 같이 소

리정보와 함께 팝업 알림을 띄운다. 사용자는 알림을 통해 그림 8 과 같은 홈 뷰를 만날 수 있다. 소리의 통계를 위해 사용자의 피드백을 서버에게 보내준다.

그림 7 는 일일/주간/월간에 대한 소리 통계 뷰이다. 한눈에 파악할 수 있으며, 개별 그래프를 클릭 시, 그 날의 혹은 그달의 세부 소리에 대한 정보를 얻을 수 있다.



(그림 6) 팝업 알림뷰

(그림 7) 통계뷰

(그림 8) 홈 뷰

그림 9 는 본 논문의 애플리케이션을 위한 녹음 프로그램 구현이다. 녹음 프로그램은 넓은 범위의 실내 소리 수집을 위하여 무지향성 마이크를 연결한 PC 에서 실행한다. 프로그램 실행 시 타이틀이 나오며, 로그인을 통해 사용자 정보를 얻는다. 로그인 성공 시 5 초 단위로 녹음이 실행되며 녹음의 시작과 끝은 'recording'과 'done recording'으로 알 수 있다. 이미지 전처리가 끝나면 'complete!' 메시지를 보여주고 추론 결과를 텍스트로 보여준다.

(그림 9) Soundee Recorder 실행화면

## 4. 결론

본 논문에서는 청각 장애인을 위한 가정에서 발생하는 소리를 딥러닝으로 예측하여 알려주는 시스템을 애플리케이션 형태로 구현하였다. 대부분 애플리케이션은 비장애인의 편리성에 맞춰진 것을 고려해 청각 장애인을 위한 서비스를 제공하였다. 본 서비스를 통해 사용자는 현재 주변의 소리를 인지할 수 있어 생활 소리를 듣지 못해 겪었던 불편함을 해소할 수 있다. 또한, 청각 장애인이 소리를 인지하지 못해 발생하는 생활 자원의 낭비를 줄이고 위험 상황 시 신속한 대처를 가능하게 할 것이라 예상한다.

본 논문으로 사회가 생각해보지 못한 청각 장애인의 일상생활 속 불편함에 대한 문제를 제기하여 이에 관한 관심과 해결방안의 마련을 도모한다. 더불어 앞으로의 4 차 산업 기술의 발전이 비장애인의 편리함뿐만이 아닌, 장애인의 접근성을 고려해야 한다는 사회적 인식을 확산시키는 데 도움이 될 것이라 예상한다.

본 논문은 과학기술정보통신부 정보통신창의 인재양성사업의 지원을 통해 수행한 ICT 멘토링 프로젝트 결과물입니다.

## 참고문헌

[1] 이슬기. (2018. 09. 27). 소소해서 더 서글픈 시청각 장애인의 불편. 에이블뉴스.

[2] 김경식. (2019. 12. 04). 청각 장애인 위험감지 · 경고 시스템 특성화 필요성-②. 에이블뉴스.

[3] 박혜연. (2020. 06. 08). “불이야” 소리쳐도… 못 피하는 ‘청각 장애인’. 충청투데이

[4] DCASE 2018 task5 Sound event detection in real life audio, http://dcase.community/challenge2018/task-monitoring-domestic-activities-results

[5] Kaggle Environmental Sound Classification 50, https://www.kaggle.com/mmoreaux/environmental-sound-classification-50

[6] Freesound 무료 저작권 사이트, https://freesound.org/

[7] 서상원, 실생활 음향 데이터 기반 이중 CNN 구조를 특징으로 하는 음향 이벤트 인식 알고리즘, 방송공학회논문지, 23, 6, 855-865, 2018