음성 특징 추출 및 딥러닝 기반 딥보이스 탐지 시스템

한승우, 한성훈, 유성민, 송동호, 서창진
상명대학교

Deep voice detection system based on voice feature extraction and deep learning

Seung-Woo Han, Seong-Hun Han, Seong-Min You, Dong-Ho Song, Chang-Jin Seo
Sangmyung University

Abstract

딥보이스는 딥러닝과 음성 합성 기술을 결합하여 생성된 가짜 음성을 의미한다. 최근 음성 합성 기술의 발전으로 딥보이스 기술은 다양한 분야에서 주목을 받고 있다. 그러나 이러한 기술이 악용될 경우 특정인 사칭, 보이스 피싱 등 여러 범죄 행위를 유발할 수 있다. 이에 본 논문은 딥보이스를 탐지해 주는 시스템을 제안한다.

본 논문에서는 MFCC와 Mel-Spectrogram을 활용하여 딥보이스와 일반 음성 데이터 셋에서 음성 특징을 추출하였다. 추출된 특징은 CNN 모델과 BiLSTM 모델을 통해 각각 학습시키고, 앙상블 기법 중 소프트보팅 기법을 사용하여 정확도 92.27% 모델을 생성하였다. 딥보이스 탐지 시스템은 사용자가 음성을 녹음하여 서버로 전송하면 앙상블 모델 기반으로 딥보이스를 탐지하고 결과를 제공한다.

본 논문에서 제안한 딥보이스 탐지 시스템은 금융 기관, 통화 기반의 고객 서비스 분야 등 딥보이스 기반 범죄가 일어날 수 있는 분야에서 안정성과 신뢰성을 향상할 수 있을 것으로 기대한다.

1. 서 론

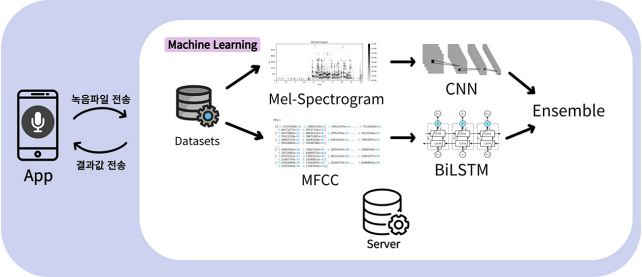
보이스피싱은 2006년 5월 국세청에 납부한 세금을 환급해 준다는 수법으로 국내에 최초로 발생한 이후에, 현재 우리 주변에서 국민 대부분이 경험하는 가장 흔한 범죄가 되었다. 그 수법과 기술이 나날이 변화함에 따라 보이스피싱 피해가 끊이지 않고 있다. 2021년 기준, 매일 80여 건 이상 발생하고 있으며, 1건당 피해 금액 2,500만 원으로 최고액을 기록하였다[1]. 또한 최근에는 딥보이스라고 알려진 딥러닝 기술이 보이스피싱 범죄자들에게 새로운 차원의 위협을 제공하고 있다. 딥보이스는 딥 러닝 알고리즘을 사용하여 특정인의 목소리를 학습하고, 그 목소리를 손쉽게 재현할 수 있다. 가족이나 지인 사칭을 포함한 보이스피싱 사례에 딥보이스가 활용되면 상당한 피해가 예상되며 해외에서는 딥보이스를 활용한 보이스피싱 사례가 실제로 보도되고 있다. 딥보이스 기술의 발전에 맞춰 보안 기술도 발전해야 하지만 아직 이와 관련된 선행 연구가 많지 않다.

따라서 본 논문에서는 딥보이스와 일반 음성을 식별해 내기 위한 딥보이스 탐지 시스템을 제안한다. 음성을 두 가지의 음성 특징 추출 기법 MFCC(Mel-Frequency Cepstral Coefficient)와 Mel-Spectrogram(Mel scale Spectrogram)을 사용하여 수치 데이터와 이미지 데이터로 변환한 후, 각각 CNN(Convolutional neural network)과 BiLSTM(Bidirectional Long Short Term Memory)으로 학습시키고, 두 알고리즘의 장점을 더 높이기 위해 소프트 보팅 앙상블 기법을 사용한다. 마지막으로 서버와 애플리케이션 사이의 통신을 통해 결괏값을 애플리케이션 화면에 출력한다.

2. 본 론

2.1 개념도 및 데이터셋

본 논문에서 제안하는 딥보이스 탐지 시스템의 개념도는 그림 1과 같다. 음성특징 추출은 MFCC, Mel-Spectrogram 이 두 가지를 사용하여 추출하였다. 데이터 셋은 AI-hub에서 제공하는 다화자 음성합성 데이터와 자유대화 음성(일반남여)데이터를 사용하였다. 이 두 가지 데이터셋 중 딥러닝 모델의 과적합 방지를 위해 다양한 화자로 딥보이스 약 25,000개, 일반 음성 약 25,000개의 음성데이터로 데이터셋을 구성하였다.

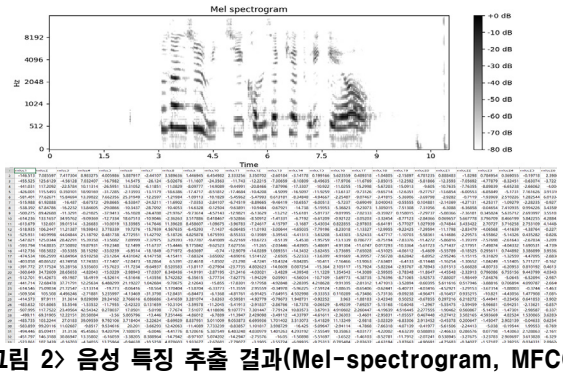


<그림 1> 제안하는 딥보이스 탐지 시스템 구성도

2.2 음성 특징 추출

딥러닝에서는 음성 자체만으로 학습이 불가능하므로 음성 특징 추출 기법을 사용해야 한다. 대표적인 음성 특징 추출 기법에는 Mel-spectrogram과 MFCC가 있다. Mel-Spectrogram은 사람의 음성 주파수 정보를 표현한 그래프이고 MFCC는 음성 신호에서 추출된 주파수 특징을 수치화하여 나타내는 계수이다. 따라서 Mel-Spectrogram을 이용하여 음성 특징을 그래프 이미지로 추출하였고 MFCC로는 수치 데이터를 추출하였다.

MFCC에서는 총 100가지의 특징을 추출하였고, 샘플링 주파수는 거의 모든 사람 음성의 범위로 16,000Hz로 설정하였다. n_fft값은 400으로 설정하였고, hop_length 값은 160으로 설정하였다. 다음으로 Mel-Spectrogram에서는 동일한 형식을 최대한 갖추기 위하여 MFCC에서 설정한 값과 같이 샘플링 주파수는 16,000Hz로, n_fft 값은 400으로, hop_length는 160으로 설정하였다. 변환할 Mel 주파수 척도는 128로 설정하였다. 그림 2는 음성특징을 추출한 결과이며, 이 결과들을 가지고 학습을 시도하였다.



<그림 2> 음성 특징 추출 결과(Mel-spectrogram, MFCC)

2.3 모델 생성

본 논문에서는 추출한 음성 특징을 CNN과 BiLSTM모델로 훈련해 모델을 생성한 뒤, 소프트 보팅기법을 이용하여 이 두 가지 모델을 앙상블한 최종 모델을 생성한다. Mel-spectrogram은 주파수 영역에서의 음성 데이터의 강도를 시간 축으로 표현한 것으로, 2D CNN 모델이 학습시키기에 가장 적합했다. 2D CNN 모델 중 주파수 영역에서의 특징을 가장 잘 추출하는 VGG19 모델을 사용하였다[2]. 음성/오디오 분석은 영상 처리와 다르게 원데이터의 형식이 일반적으로 1차원 데이터라는 점과, 시계열이라는 특징을 가지고 있다[3]. RNN을 사용하면 이러한 시계열 정보를 활용하여 음성의 특징과 구조를 잘 파악할 수 있다. 다양한 RNN 모델 중 현재 가장 성능이 좋은 BiLSTM을 사용하였다.

먼저, CNN 모델은 tensorflow에서 제공하는 VGG19 모델을 사용하였다. 또한, 사전 학습된 가중치를 사용하였으며, 최상위 Layer는 포함하지 않았다. VGG19모델 뒤에 Flatten layer와 Dense layer, 과적합 방지를 위하여 Dropout layer를 사용하였다. 사용된 모델의 Dropout layer는 0.5이고 512개의 뉴런을 가진 Dense layer에 활성화 함수는 ReLu를 2개의 뉴런을 가진 Dense layer를 이어 붙였으며, Adam optimizer를 사용하였다. 다음으로 BiLSTM 모델의 설정값은 sequence model이고, units 수는 512, Dropout layer는 0.8이다. 그리고 sigmoid 활성화 함수와 Adam optimizer를 사용하였다. 끝으로, 소프트 보팅은 각각 학습된 모델들이 카테고리별로 가능성을 가지고 평균을 내어 가장 높은 확률의 카테고리를 투표로 선정하는 알고리즘이다[4]. 제안한 모델은 두 가지 모델을 사용하기 때문에 앙상블 기법 중 소프트 보팅 기법을 선정하였다.

2.3.1 실험결과

개발 및 실험에 사용한 환경은 다음과 같다. OS는 Windows 11 Home을 사용하였고, CPU는 Ryzen 9 5900X이며, GPU는 NVIDIA GeForce TRX 3070에 RAM은 32GB이다. 그리고 딥러닝 알고리즘 수행을 위해 Tensorflow, Python 3, Sklearn을, 음성 특징 추출을 위해 Librosa를 사용하였다. VGG19모델과 BiLSTM 모델을 앙상블 하였다. 테스트 셋은 학습데이터와 전혀 관련이 없는 직접 녹음한 데이터셋 329개와 클로바를 통해 생성한 딥보이스 370개를 가지고 진행하였다. 표 1을 보면 VGG19의 테스트 정확도는 91.27%가 나왔으며, BiLSTM의 테스트 정확도는 80.97%가 나왔다. 이 두 가지 모델을 앙상블 하였을 때의 테스트 정확도는 92.27%가 나왔다. 실제로 사용 가능할 정도의 정확도가 나온다고 판단하여 생성된 모델을 애플리케이션에 결합하였다.

<표 1> 모델들의 이진분류 성능평가

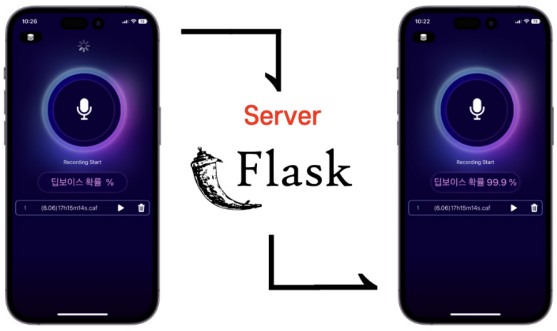
Algorithm	Precision	Recall	F1-score	Accuracy
CNN	87.43	95.13	91.11	91.27
BILSTM	89.36	75.00	81.55	80.97
CNN+BILSTM Ensemble	91.54	92.09	91.81	92.27

2.4 애플리케이션

본 애플리케이션은 리액트 네이티브 기술을 활용하여 개발하였다. 리액트 네이티브는 자바스크립트 기반의 모바일 애플리케이션 개발 프레임워크이다. 이를 통해 iOS와 안드로이드 플랫폼에서 모두 동작하는 애플리케이션을 효율적으로 작성할 수 있으며, 개발 시간을 단축하고 효율성을 높일 수 있다.

본 서버는 파이썬 웹 프레임워크인 플라스크(Flask)를 활용하여 개발하였다. 플라스크는 딥러닝 모델 및 기타 파이썬 라이브러리의 통합과 연동이 원활하기 때문에, 논문에서 제안한 시스템의 목적에 부합하는 기능을 제공한다.

사용자는 애플리케이션을 실행하고 딥보이스로 의심되는 목소리를 녹음하여 서버로 전송한다. 서버에서는 구축된 모델을 기반으로 전송된 음성의 딥보이스를 탐지한다. 탐지 결과는 그림 3과 같이 애플리케이션 화면에 시각적으로 출력되어 사용자에게 제공한다.



<그림 3> 녹음 후 서버와 통신하여 결과 값을 받아오는 장면

3. 결 론

본 논문은 딥보이스 탐지 시스템을 제안하였다. AI-hub에서 제공한 데이터 중 딥보이스 약 25000개, 일반 음성 약 25000개를 선정하여 음성 특징 추출 기술 MFCC, Mel-Spectrogram을 이용해 특징을 추출하였다. 이를 바탕으로 CNN-VGG 19 모델과 BiLSTM 모델을 사용해 각각 테스트 정확도 91.27%, 80.97%의 모델을 만들었으며 이 두 개의 모델을 앙상블 하여 최종적으로 정확도 92.27%의 모델을 만들었다. 애플리케이션과 서버는 각각 리액트 네이티브 기술과 플라스크를 활용하여 구현했다. 이를 통해 사용자는 딥보이스로 의심되는 목소리를 녹음하여 서버로 전송하고, 서버는 딥보이스를 식별한 후 결과를 시각적으로 제공한다.

본 논문에서 제안된 딥보이스 탐지 시스템은 금융 기관 및 통화 기반의 고객 서비스 분야 등 다양한 분야에서 딥보이스 기반 범죄 예방에 기여할 수 있으며, 통화상에서의 안정성과 신뢰성을 향상할 수 있다. 또한, 향후 연구에서는 다양한 환경에서 음성 데이터를 수집하고 해당 시스템을 발전시킴으로써 딥보이스에 대한 탐지 정확도를 더욱 향상할 것으로 예상된다. 이러한 연구의 발전은 사회적으로 딥보이스 관련 범죄 피해를 근절하는 데 크게 기여할 것으로 기대된다.

[참 고 문 헌]

[1] 서준배, "[범죄·안전] 보이스피싱 현황, 유형, 추이와 대응관련 시사점" 온라인.https://kostat.go.kr/board.es?mid=a90104010311&bid=12312&tag=&act=view&list_no=422196&ref_bid=.(2023.04.05.)

[2] R. Reimao, "SYNTHETIC SPEECH DETECTION USING DEEP NEURAL NETWORKS," Master's thesis, York University, Toronto, Ontario, Canada, May 2019.

[3] 이영한, "딥러닝 기반의 음성/오디오 기술", 방송과 미디어, 제22권 제1호, 46-57, (2017)

[4] 전병욱, 강지수, 정경용. "도로교통 이머징 리스크 탐지를 위한 AutoML과 CNN 기반 소프트 보팅 앙상블 분류 모델" 융합정보논문지, 11(7), 14-20. (2021)