ISSN 1229-800X [Print] / ISSN 2586-7792 [Online]
The Transactions of the Korean Institute of Electrical Engineers, vol. 72P, no. 3, pp. 186~192, 2023
https://doi.org/10.5370/KIEEP.2023.72.3.186

A Study on the Development of Deep Learning-Based Deep Voice Detection System Using Mel-Spectrogram and MFCC

Mel-Spectrogram과 MFCC를 이용한 딥러닝 기반 딥보이스 탐지시스템 개발에 관한 연구

Seung-Woo Han · Seong-Hun Han · Seong-Min You · Dong-Ho Song · Chang-Jin Seo
한승우* · 한성훈* · 유성민* · 송동호* · 서창진†

Abstract

Deep voice refers to a fake voice produced with deep learning and voice synthesis technology. In this paper, we propose a deep-learning-based deep voice detection system using MFCC and Mel-Spectrogram. We propose an ensemble model using CNN (Convolution Neural Network) and BiLSTM for the development of deep voice detection systems. In the experiment, the training dataset used voice data provided by AI-HUB about 50,000 voice data, 25,000 each for the deep voice and general voice. And the test dataset was created with 370 deep voices generated from NAVER CLOVA and 329 directly recorded datasets. And a 92.27% accuracy model was constructed using the soft-voating method. The deep voice detection system detects deep voices based on the ensemble model and provides results when the user records the voice and transmits it to the server. The deep voice detection system proposed in this paper is expected to improve stability and reliability in areas where deep voice-based crime.

Key Words

Deep Voice, Mel-Spectrogram, Bi-LSTM, CNN, MFCC, Voice Synthesis

1. 서 론

보이스피싱은 국내·외에서 급속하게 증가하고 있는 범죄 중 하나이다. 2006년 5월 국세청에 납부한 세금을 환수해 준다는 수법을 시작으로 현재까지 꾸준히 발생하고 있으며 국민의 경험 중 가장 흔한 범죄 사례 중 하나로 자리매김하고 있다. 이에 따라 보이스피싱의 수법과 기술도 변화하고 발전하고 있다. 2022년 통계청에서 제공하는 보이스 피싱 관련 통계 자료에 의하면, 2006년 한해에 1,488건의 보이스피싱 범죄가 발생하였으며, 이후 꾸준히 증가하여 2018년에는 3만 4,132건이 발생해 최초로 3만 건을 돌파한 후, 2019년에는 3만 7,667건으로 한 해 발생 건수로는 최대를 기록하였다. 2021년 기준, 매일 80건 이상 발생하고 있으며, 1건당 피해 금액 2,500만 원으로 최고액을 기록하였다[1]. 더욱이 최근에는 딥러닝을 기반으로 한 음성 조작 기술인 '딥보이스'가 등장하여 보이스피싱의 위협이 더욱 증가하고 있다. 이러한 딥보이스 기술은 딥러닝 알고리즘을 활용하여 특정 개인의 목소리를 학습하고, 그 목소리를 손쉽게 재현할 수 있는 기술이다. 이러한 음성 조작 기술의 발전으로 누구나 손쉽게 음성을 조작할 수 있게 되었다. 기술의 발전으로 인해서 누구나 손쉽게 음성 조작이 가능해진 만큼 앞으로 다양한 유형의 범죄가 증가할 가능성도 커질 것으로 예측할 수 있을 것이다[2]. 이러한 딥보이스 기술을 이용하여 가족이나 지인을 사칭한 보이스피싱 사례가 더욱 늘어날 것으로 예상되며, 해외에서는 딥보이스를 활용한 보이스피싱 사례가 실제로 보도되고 있다. 이렇듯 딥보이스 기술의 발전함에 따라 보안 기술의 발전의 필요성이 점점 커지고 있다. 이에 따라 최근, 딥보이스와 일반 음성의 특징을 비교하여 수치로 나타낸 연구가 있었으며, 또한 화자식별 및 화자검증 기술을 사용하여 딥보이스와 일반 음성을 구별해주는 서비스에 대한 연구도 찾을 수 있었지만, 아직까지 딥보이스에 대한 관련 연구가 많이 부족한 실정이다.

따라서 본 논문은 딥러닝 기반 딥보이스 탐지 시스템을 제안한다. 제안하는 연구는 AI-HUB에서 제공하는 음성 데이터를 활용하였다. 딥보이스에는 다화자 음성 합성 데이터를 사용하였고, 일반 음성에는 자유 대화 음성 데이터를 사용하여 각 25,000개씩 약 5만 개의 음성 데이터를 수집하였다. 이렇게 수집한 음성 데이터를 MFCC (Mel Frequency Cepstral Coefficient)와 Mel-Spectrogram이라는 두 가지 음성 특징 추출 기법을 사

† Corresponding Author : Dept. of Information Security Engineering Sangmyung University, Korea.
E-mail: cjseo@smu.ac.kr https://orcid.org/0000-0002-9874-4950
* Dept. of Information Security Engineering Sangmyung University, Korea.
https://orcid.org/0009-0000-3523-9781 https://orcid.org/0009-0008-4381-1248
https://orcid.org/0009-0000-1768-8727 https://orcid.org/0009-0001-1743-3166
Received : Aug. 11, 2023 **Revised :** Aug. 22, 2023 **Accepted :** Aug. 28, 2023

용하여 수치데이터와 이미지 데이터로 변환한다. 변환된 데이터는 CNN (Convolutional Neural Network)과 BiLSTM (Bidirectional Long Short-Term Memory)이라는 딥러닝 알고리즘을 사용하여 학습되며, 이 두 알고리즘의 장점을 극대화하기 위해 소프트 보팅 앙상블 기법을 도입하여 최종 모델을 구축하였다. 마지막으로 구축한 모델을 사용자가 쉽게 활용할 수 있도록 서버와 연결된 애플리케이션을 개발하였다.

2. 제안하는 시스템 구성

2.1 데이터 세트

본 논문에서 제안하는 딥보이스 탐지 시스템의 개념도는 그림 1과 같다. 음성 특징 추출은 MFCC, Mel-Spectrogram을 사용하여 추출하였다. 데이터 세트는 AI-HUB에서 제공하는 다화자 음성 합성 데이터와 자유 대화 음성(일반 남녀) 데이터를 사용하였다. 일반 음성 데이터 세트는 10대에서 50대 사이의 일반인 남녀의 발화 데이터로 총 2,000명 이상, 약 4,000시간 음성 데이터로 구성되어 있으며, 성별 분포는 남녀 1:1 비율이다. 딥보이스 데이터 세트는 10대에서 60대까지 일반인 남녀 3,400명 이상, 약 10,000시간 음성 데이터로 만들었으며, 성별 분포는 남녀 2:3 비율이다. 딥보이스 모델로는 Tacorton2를 사용하였다. 두 가지 데이터 세트 중 딥러닝 모델의 과적합 방지를 위해 다양한 화자로 딥보이스 약 25,000개, 일반 음성 약 25,000개의 음성 데이터로 데이터 세트를 구성하였다.

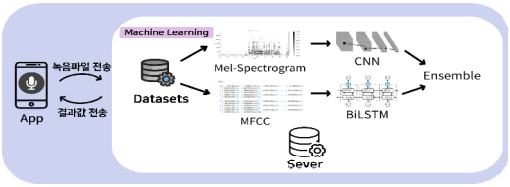


그림 1 제안하는 딥보이스 탐지 시스템 구성도
Fig. 1 The Proposed System Configuration

2.2 음성 특징 데이터 구축

딥러닝에서는 음성 자체만으로 학습할 수 없으므로 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 MFCC를 이용하여 수치데이터를 추출하였다.

2.2.1 Mel-Spectrogram

Mel-spectrogram은 SFTF (Short Time Fourier Transform)을 바탕으로 한 Spectrogram 성분을 인간의 청각 특성을 잘 나타낼 수 있도록 Mel Scale로 변환한 것이다. Mel Scale은 인간의 청각 특성을 구조화하여 1 kHz 이하에서는 중심 주파수가 균일하게 나누어지고 1 kHz 이상에서는 Logarithm Scale에 따른다[3]. 변환된 멜 스케일 스펙트럼을 이미지화한 것이 멜 스펙트로그램이며, 시간에 따라 진폭 축의 변화를 시각적으로 볼 수 있는 파형과 주파수 상에서 진폭 축의 변화를 볼 수 있는 스펙트럼의 특징이 결합 되어 있으며, 진폭의 차이를 색으로 나타낸다[4]. Mel-Spectrogram은 Mel Filter Bank가 적용된 Spectrogram으로, 각각의 프레임에 대해 얻어낸 주파수들에 대해서 Mel 값을 얻어내기 위한 Filter를 적용한다. 또한 이 Mel Filter Bank는 주파수 Scale을 지수적 증가로 표현하는데 이는 사람이 소리를 인식하는 것과 유사하다. 그리하여 사람이 잘 인식할 수 있는 저음역의 주파수는 세밀하게 표현하며 그 외 영역은 상대적으로 덜 세밀하게 표현한다[5]. 따라서 본 논문에서 다루고자 하는 딥보이스와 일반 음성 데이터의 특징을 추출하는 데 있어서 적합한 기법이라고 할 수 있다. 그림 2는 가공되지 않은 음성 신호와 음성 신호를 Mel-Spectrogram을 이용하여 변환한 결과이다.

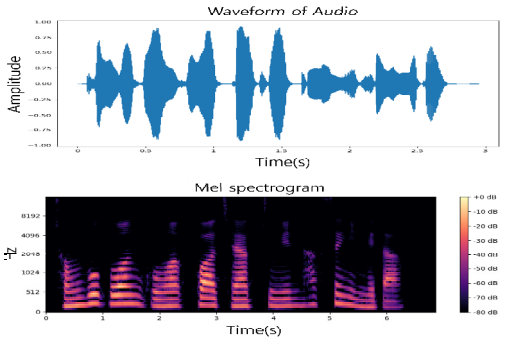


그림 2 음성 데이터 mel scale 추출 전과 추출 후
Fig. 2 Before and after extracting voice data mel scale

2.2.2 MFCC

MFCC는 음성 신호 처리 분야에서 널리 사용되는 효과적인 특징 추출 기술로, 음성 신호를 주파수 정보의 고차원 표현에서 더 의미 있는 저차원의 특징 공간으로 변환한다. 이는 음성 인식, 화자 인식, 음성 합성 등 다양한 음성 처리 작업에서 중요한 역할을 하며, 음성 신호의 주파수 정보를 사람의 청각 특성에 더 가깝게 변환하여 중요한 음성 특징을 추출하는 과정이다. MFCC 특징값을 뽑아내는 과정은 그림 3과 같다.

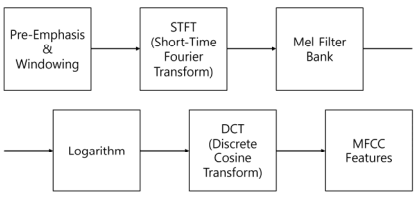
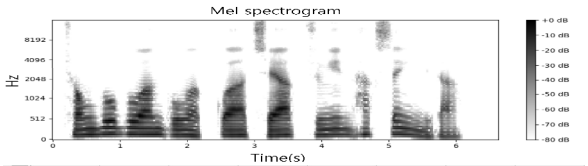


그림 3 MFCC 특징 추출 과정
Fig. 3 MFCC feature extraction process

초기 단계로 입력 신호에 Pre-Emphasis Filter를 적용하여 고주파 신호의 에너지를 증가시키고, 이에 따라 고주파 잡음의 영향을 최소화하는 효과를 얻는다. 다음으로, 단시간(Short-Time) 푸리에 변환을 수행하여 시간 도메인의 데이터를 주파수 도메인으로 변환한다. 이러한 단시간 푸리에 변환은 신호를 작은 구간으로 나누어 각 구간에서의 주파수 정보를 분석하며, 이를 통해 시간에 따른 주파수 변화를 Spectrogram으로 확인할 수 있다. 변환 구간의 크기와 오버랩을 조절하는 윈도우(Window)와 홉(Hop)의 크기를 설정하여 변환의 정밀도를 조절한다. 이후, Mel Filter Bank를 적용하여 주파수 스펙트럼을 Mel Scale로 변환한다. Mel Filter Bank는 앞서 언급한 것과 같이, 일련의 Mel Scale Filter를 통해 신호의 주파수를 사람의 청각 특성에 맞게 변환하며, 이를 통해 선형적이지 않은 주파수 영역을 보다 적절한 비선형 스케일로 변환한다. 변환된 주파수 정보를 이용하여 Mel Scale에 근사시키며, 이 과정에서 로그 변환을 통해 주파수의 비선형적 특성을 고려한다. 변환된 데이터에 이산 코사인 변환(DCT)을 적용하여 MFCC 특징값을 추출하며, 일반적으로 13개의 계수를 사용한다. 또한 음성 인식 분야에서는 Delta 값과 Delta-Delta 값을 추가적인 특징값으로 이용하여 잡음에 강인한 모델을 구성하고 성능을 향상시킬 수 있다[6, 7].

2.2.3 음성 특징 추출

제안하는 연구에서는 MFCC를 이용하여 100가지의 음성 특징을 추출하였고, 샘플링 주파수는 16,000Hz로 설정하였다. 음성은 연속적인 신호이며, 이런 연속적인 신호를 주파수로 바꿔야 하는데, 이때 사용하는 것이 n_fft 이며, 값이 높아질수록 높은 주파수의 해상도를 가진다. n_fft 값은 400으로 설정하였고, hop_length 값은 음성을 작은 조각으로 자를 때, 자르는 간격을 의미하며, 160으로 설정하였다. 이는 샘플링 주파수로 설정한 16,000Hz에 대한 균형을 맞춘값으로, 중간 크기의 주파수와 작은 간격으로 프레임을 생성해 시간적 변화를 높여줄 수 있다. 다음으로 Mel-Spectrogram에서는 같은 형식을 최대한 갖추기 위하여 MFCC에서 설정한 값과 같이 샘플링 주파수는 16,000Hz로, n_fft 값은 400으로, hop_length는 160으로 설정하였다. Mel 주파수 척도는 128로 설정하였다. 마지막으로 추출된 Mel-spectrogram에 흑백화를 적용하여 학습 정확도를 향상시켰다. 그림 4는 Mel-Spectrogrm과 MFCC를 이용한 음성 특징을 추출한 결과이며 학습에 적용하였다.



	mfcc1	mfcc2	mfcc3	mfcc4	mfcc5	mfcc6	mfcc7	mfcc8
2	-484.3022	81.11156	19.06075	27.41708	-9.984256	3.545862	-7.443961	-4.067462
3	-479.0167	71.86276	24.92076	32.31567	-14.68362	5.074169	-9.709321	-6.509709
4	-482.4822	66.85391	24.58688	21.77356	-13.06513	9.953751	-7.658546	-5.133058
5	-486.5508	83.02607	17.79838	28.0113	-13.67127	3.66475	-8.506121	-4.635557
6	-474.791	90.76698	18.00949	10.70524	-11.35785	12.36617	-13.36567	-4.262253
7	-475.3019	73.94495	24.60814	26.75244	-8.503124	5.156363	-5.852896	-8.159203

그림 4 음성 특징 추출 결과(Mel-spectrogram, MFCC)
Fig. 4 Voice feature extraction results

2.3 모델 생성

본 논문은 음성 데이터에서 추출한 특징을 활용하여 CNN과 BiLSTM 모델을 훈련하고, 이 모델들을 소프트 보팅 기법을 활용하여 효과적으로 앙상블하는 방법에 초점을 맞추고 있다. 특히 Mel-Spectrogram은 음성 데이터의 주파수 영역에서 강도를 시간 축으로 표현한 것으로, 이러한 형태의 데이터는 2D CNN 모델의 학습에 가장 적합하다. 본 논문에서는 2D CNN 모델 중에서 주파수 영역에서의 특징을 가장 잘 추출하는 VGG 19 모델을 선택하여 사용하였다 [8]. 음성 및 오디오 데이터 분석은 일반적인 영상 처리와 다르게 1차원 데이터의 형태를 가지며, 이러한 데이터의 시계열 특성이 뚜렷하다 [9]. 이에 따라 RNN을 활용하면 시계열 정보를 적극적으로 활용하여 음성 데이터의 특징과 구조를 잘 파악할 수 있다. 다양한 RNN 모델 중에서 성능이 뛰어난 BiLSTM을 선택하여 사용하였다. 이와 같은 과정을 통해 음성 데이터의 복잡한 특성을 최대한 활용하여 앙상블 모델을 구성 하였다.

2.3.1 VGG-19

CNN 모델은 TensorFlow의 제공하는 VGG-19 모델을 활용하여 실험을 진행한다. VGG 19는 19개의 계층으로 이루어진 CNN 구조로, 주로 이미지 분류를 위한 목적으로 설계되었다. 그림 5에서 확인할 수 있듯이, VGG 19의 각 계층은 컨볼루션(Convolution), 활성화 함수, 풀링(Pooling) 계층으로 이루어져 있으며, 총 16개의 컨볼루션 계층과 3개의 완전 연결 계층으로 구성되어 총 19개의 계층을 가지고 있다[10]. 계층 구조는 VGG 19가 이미지의 다양한 특징을 추출하고 학습하는 강점을 제공한다. 특히 3x3 크기의 작은 커널을 여러 번 중첩하여 컨볼루션 계층을 구성함으로써 모델은 더 복잡하고 추상적인 패턴을 학습하며 동시에 파라미터 수를 줄여 메모리와 계산 비용을 효율적으로 관리할 수 있다. 실험에서는 사전 학습된 가중치를 사용하여 모델의 초기 가중치를 설정하였다. 최상위 Layer는 제외하였으며, VGG 19 모델 이후에는 Flatten Layer를 통해 데이터를 일차원으로 변환하고, 이어서 Dense Layer와 Dropout Layer가 사용되었다. Dropout Layer는 과적합을 방지하기 위해 활성화 함수를 통해 얻은 중간 출력값을 무작위로 삭제하는 역활을 수행한다. 이 모델은 0.5의 Dropout 비율을 가지며, 512개의 뉴런을 가진 Dense Layer와 ReLU 활성화 함수를 통해 복잡한 비선형성을 학습하였다. 마지막으로, 2개의 뉴런을 가진 Dense Layer를 추가하여 최종 클래스 분류를 수행하였으며, Adam Optimizer를 사용하여 모델의 가중치를 조정하였다.

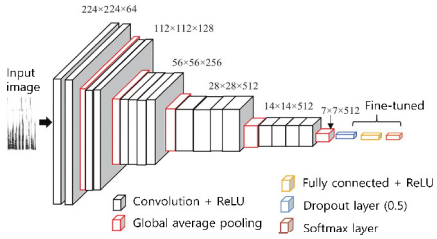


그림 5 VGG 19 구성도
Fig. 5 VGG 19 Configuration

2.3.2 BiLSTM

BiLSTM은 순차적인 데이터 처리에 있어서 LSTM의 발전된 형태로서, 시계열 데이터나 음성 데이터와 같이 시간적인 의존성을 갖는 데이터를 다루는 데 매우 유용한 역할을 한다. LSTM은 기본적으로 과거의 정보를 기억하고 현재의 입력을 기반으로 다음 상태를 예측하는 데 사용된다. 그러나 RNN이나 LSTM은 입력한 순서를 시간 순서대로 입력하기 때문에 결과물이 직전 패턴을 기반하여 수렴하는 경향이 있다는 한계가 보인다[11]. 이에 반해 BiLSTM은 양방향으로 데이터를 처리하는 구조로 되어 있어 과거와 미래의 정보를 모두 고려할 수 있다. 그림 6과 같이 두 방향의 LSTM 결과를 결합함으로써, 각 시점에서 과거와 미래의 정보를 함께 고려하면서 더욱 정확한 패턴 인식을 가능케 한다. 음성 데이터의 경우, 발화자의 발음 특징이 시간상으로 변하는 특성 때문에 시계열적인 의미가 크다. 따라서 BiLSTM은 음성 데이터에서 음절, 발음, 강세 등의 특징을 더욱 정확하게 추출하고 이를 기반으로 판별하는 데 도움을 줄 수 있다. 특히 딥보이스와 정상 음성을 구분하는 작업에서 BiLSTM은 양 방향적인 정보 처리를 통해 딥보이스의 특징을 탐지하고 이를 구별하는 데 유용하게 활용된다. 따라서, 본 연구에서는 음성 데이터의 시계열적인 특성을 고려하고 BiLSTM을 활용하여 딥보이스와 정상 음성을 식별하고 탐지하는 시스템을 개발하였다. 이를 통해 더 정확하고 신뢰성 있는 딥보이스 탐지를 실현하며, 음성 데이터의 복잡한 패턴을 잘 인식하는 데 기여했다. BiLSTM 모델의 설정값으로는 Sequence Model이고, Units 수는 512, Dropout Layer는 0.8이다. 그리고 Sigmoid 활성화 함수와 Adam Optimizer를 사용하였다.

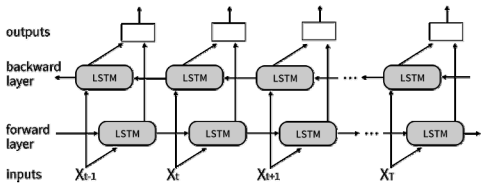


그림 6 BiLSTM 구성도
Fig. 6 BiLSTM Configuration

2.3.3 앙상블(Ensemble) 모델

본 논문에서 제안하는 최종 모델인 CNN+BiLSTM 앙상블 모델을 만들기 위하여, 각각의 모델을 생성한 뒤, 앙상블을 진행하였다. 먼저, 앙상블이란 크게 부스팅(Boosting), 배깅(Bagging), 보팅(Voting) 등으로 나누어진다. 그중 보팅 기법은 여러 모델을 통해 얻은 예측값들을 대상으로 투표하는 방식으로 최종 예측값을 정하는 것이다. 보팅 기법에는 하드 보팅과 소프트보팅이 있다. 하드 보팅은 여러 모델의 클래스 예측값들을 가지고 다수결 투표를 하여 가장 많이 나온 클래스를 최종 예측값으로 지정하는 방식이다. 소프트 보팅은 그림 7과 같이 각 모델의 예측값들을 가지고 평균을 내어 최종 클래스를 최종 예측값으로 지정하는 방식이다. 하드 보팅의 경우 신뢰성을 위해서는 많은 모델이 필요하다. 만일, 모델이 적은 상황에서 오판단하거나 예측값이 높지 않을 때는 최종 클래스 예측 자체가 오판단일 가능성이 높다. 하지만 소프트 보팅의 경우에는 적은 모델의 수를 가지고도 보다 높은 정확도를 보일 수 있으며, 오판단이나 예측값이 높지 않을 때도 각 클래스의 평균을 내어 최종 클래스를 예측하기 때문에 신뢰성과 정확도를 보다 높일 수 있다[12]. 본 논문에서 제안한 모델은 두 가지 모델을 사용한다. 그렇기 때문에 하드 보팅에는 적합하지 않고, 소프트 보팅이 적합하며, 더 높은 정확도를 보일 수 있기에 앙상블 기법 중 소프트 보팅 기법을 선정하였다.

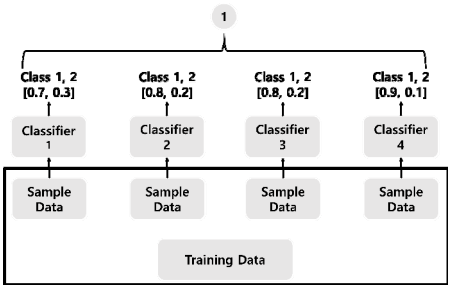


그림 7 소프트 보팅 기법
Fig. 7 Soft-Boating Method

3. 실험 및 성능평가

3.1 성능평가

개발 및 실험에 사용한 환경은 다음 표 1과 같으며, 딥러닝 알고리즘 수행을 위해 Tensorflow, Python 3, Sklearn을, 음성 특징 추출을 위해 Librosa를 사용하였다.

표 1 시스템 개발환경
Table 1 System Development Environment

OS	- Windows 11 Home 64bit
CPU	- Ryzen 9 5900X
GPU	- NVIDIA GeForce RTX 3070
RAM	32GB

본 논문에서는 성능평가를 위해서이진 분류 (Binary Classification)

작업을 수행하였다. 또한 이진 분류 작업에서 모델의 성능을 평가하고 예측 결과를 분석하는 데 많이 사용되는 혼동 행렬 (Confusion Matrix)을 사용하였다. 혼동 행렬은 실제 클래스와 모델의 예측 결과 간의 관계를 나타내며, 표 2에서 TP (True Positive), FP (False Positive), TN (True Negative), FN (False Negative) 4개의 항목을 확인할 수 있다. 이를 활용하여 다양한 성능평가 지표를 계산할 수 있다.

3.2 성능평가 지표

성능평가에서는 정밀도 (Precision), 재현율 (Recall), F1-Score, 정확도 (Accuracy) 지표를 계산하였다. 정밀도는 모델이 양성으로 예측한 샘플 중에서 실제 양성인 샘플의 비율을 나타내고 재현율은 실제 양성이 샘플 중에서 모델이 양성으로 예측한 샘플의 비율을 나타낸다. F1-Score는 정밀도와 재현율의 조화 평균으로 계산되며 모델의 양성 예측 능력을 종합적으로 평가하고 정확도는 모델이 전체 샘플 중에서 정확히 예측한 샘플의 비율을 나타낸다. 4개의 성능평가 지표를 계산하는 식은 식 (1), (2), (3), (4)와 같다.

$$\mathrm{Precision} = \frac{TP}{TP+FP} \tag{1}$$

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

$$F1\,score = \frac{2\times \mathrm{Precision}\times Recall}{\mathrm{Precision} + Recall} \tag{3}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

표 2 혼동 행렬 (Confusion Matrix)
Table 2 Confusion Matrix

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

표 3 모델들의 이진 분류 성능평가
Table 3 Binary classification performance evaluation of models

Algorithm	Precision	Recall	F1-score	Accuracy
CNN	87.43	95.13	0.9111	91.27
BiLSTM	89.36	75.00	0.8155	80.97
Ensemble	91.54	92.09	0.9181	92.27

3.3 실험 결과

본 논문에서는 VGG 19 모델과 BiLSTM 모델을 앙상블 하여 제안하는 최종 모델을 구축하였다. 테스트 세트는 학습데이터와 전혀 관련이 없는 직접 녹음한 데이터 세트 329개와 네이버 클로바를 통해 생성한 딥보이스 370개를 가지고 진행하였다. 표 3을 보면, CNN 모델의 경우 정밀도는 87.43%, 재현율은 95.13%, F1-Score는 0.9111, 그리고 정확도는 91.27%로 나타났다. 이 모델은 긍정 클래스를 정확하게 예측하는 데 높은 능력을 보이고 있으며, 전체적인 분류 정확도 또한 높은 수준이다. 다음으로 BiLSTM 모델의 경우 정밀도는 89.36%, 재현율은 75%, F1-Score는 0.8155, 그리고 정확도는 80.97%로 나타났다. 이 모델은 긍정 클래스의 예측에는 높은 정밀도를 보이지만, 재현율이 낮은 편이다. 따라서 긍정 클래스 중 일부를 놓치는 경향을 보였다. 마지막으로, CNN+BiLSTM Ensemble 모델의 경우 정밀도는 91.54%, 재현율은 92.09%, F1-Score는 0.9181, 그리고 정확도는 92.27%로 나타났다. 이 모델은 정밀도와 재현율 모두 높은 수준을 보이며, F1-Score와 정확도 역시 높은 편이다. 이는 CNN과 BiLSTM 모델 각각의 강점을 효과적으로 결합한 결과로 볼 수 있다. 따라서 실험 결과를 종합해 보면, CNN+BiLSTM Ensemble 모델이 가장 뛰어난 성능을 보이며 다양한 평가 지표에서 우수한 결과를 나타냈다. 실제로 사용할 수 있는 정도의 정확도가 나온다고 판단하여 생성된 모델을 애플리케이션에 결합하였다.

3.4 애플리케이션 및 서버 구현

본 연구에서 제안하는 딥보이스 탐지 시스템을 활용하여 애플리케이션을 작성하였다, 모바일 애플리케이션 개발 프레임워크인 리액트 네이티브와 웹 프레임워크인 Flask를 기반으로 개발되었다. 본 절에서는 애플리케이션과 서버의 구현에 관해 설명한다. 애플리케이션은 그림 8과 같이 사용자가 애플리케이션을 실행하고 녹음을 시작한다. 녹음이 완료되면, 애플리케이션은 녹음된 음성 파일을 서버로 전송한다. 서버는 받은 파일을 딥보이스 탐지 모델에 제공하여 음성을 분석하고, 딥보이스 음성을 탐지한다. 이후 서버는 탐지 결과를 애플리케이션으로 반환하고, 사용자는 화면에 나타난 결과를 통해 딥보이스 탐지 여부를 시각적으로 확인할 수 있다.

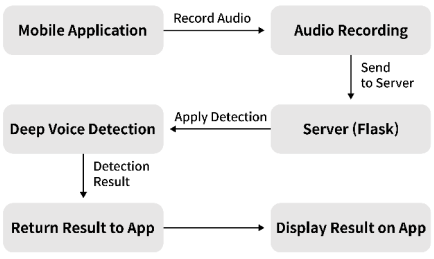


그림 8 애플리케이션 구성도
Fig. 8 Application Configuration

서버는 딥보이스 탐지 시스템의 핵심 기능을 수행하는 역할을 맡고 있다. 서버는 Flask 기반의 웹 프레임워크를 사용하여 구축되어 있으며, 간결하고 가벼운 구조로 되어 있어 작은 규모부터 중간 규모의 웹 애플리케이션을 빠르게 개발할 수 있도록 지원한다. 클라이언트는 딥보이스 음성 탐지를 수행하기 위

해 POST 메서드를 활용하여 오디오 파일을 서버로 전송한다. 서버는 전송된 오디오 파일을 받아서 특정 폴더에 저장한다. 이후, 서버는 저장된 오디오 파일을 딥보이스 탐지 모델에 적용하여 분석하고, 탐지 결과는 JSON 형식으로 클라이언트에 반환된다. 클라이언트는 이 결과를 활용하여 딥보이스 음성 탐지 결과를 애플리케이션 화면에 표시하게 된다.

이 시스템을 통해 사용자는 어디서든 간편하게 딥보이스 탐지를 수행할 수 있으며, 모바일 환경의 편의성과 접근성을 즐길 수 있다. 사용자가 일상적인 환경에서 의심스러운 전화나 영상을 마주했을 때, 모바일 애플리케이션을 실행하여 음성을 녹음하고 딥보이스 탐지 시스템을 활용하여 해당 녹음이 딥보이스 음성인지 확인할 수 있다. 이로써 사용자는 딥보이스를 이용한 범죄로부터 안전을 확보할 수 있다.

4. 결 론

보이스피싱은 국내에서 최초로 등장한 이후로 지속적으로 발전하고 흔해진 범죄 중 하나로 부상하고 있다. 특히 딥보이스 기술의 등장으로 그 위협은 더욱 커지고 있다. 이런 상황에서 본 논문은 딥보이스와 일반 음성을 식별해 내는 딥보이스 탐지 시스템을 제안하였다. 실험 결과는 상당히 좋은 성과를 보여주었으며, CNN-VGG 19 모델과 BiLSTM 모델의 개별 테스트에서도 각각 정확도 91.27%, 80.97%로 높은 정확도를 기록하였으며, 두 모델을 앙상블 하여 테스트 정확도를 92.27%의 최종 모델을 만들었다. 본 논문에서 제안한 딥보이스 탐지 시스템은 보이스피싱과 같은 범죄 예방에 높은 잠재력을 가지고 있어, 음성 통화를 통한 사기 사례를 예방하고, 금융 기관과 고객 서비스 분야에서 안정성과 신뢰성을 높이는 데 크게 기여할 것으로 기대된다. 또한, 향후 연구에서는 다양한 환경에서 음성 데이터를 수집하고 해당 시스템을 발전시킴으로써 딥보이스에 대한 탐지 정확도를 더욱 향상할 것으로 예상된다.

References

[1] JunBae Seo, “Current status, type, trend, and response implications of voice phishing,” Statistics Korea, 2022. https://kostat.go.kr/synap/skin/doc.html?fn=2f1b5a46a70ffa293ac2113cbd1b9635e28497f762079ae221b0d41b4129ca9d&rs=/synap/preview/board/12312/.

[2] Sowoon Kim, Sungtaek Lee, “Development of Voice Phishing Damage Prevention Service Misusing Deep Voice,” The Journal of Korean Institute of Communications and Information Sciences, vol 47, no 10, pp. 1677-1685, 2022. DOI : 10.7840/kics.2022.47.10.1677.

[3] Youngjun Sim, Jungyu Choi, Sungbin Im, “Synthetic Speech Classification based on Cascade Connection of CNN and MKDE Models,” Journal of the Institute of Electronics and Information Engineers, vol 60, no 2, pp. 94-101, 2023. DOI : 10.5573/ieie.2023.60.2.94.

[4] JeongKyu Oh, “Acoustic Classification using Convolutional Neural Networks : Application to Beehive Sound,” Master's thesis, Chungbuk National University, 2021.

[5] JungGuk Lee, “Deep learning based heart murmur detection using Mel Spectrogram and Peak Interval Feature of heartbeat sounds,” Master's thesis, Chungang University, 2023.

[6] HyunSub Kim, “Sleep arousal detection model based on recurrent neural networks using MFCC as a feature vector,” Master's thesis, KAIST, 2019.

[7] Kyeongsik Shin, Sinwoo Yoo, Hyukjun Oh, “Detection and Classification for Low-altitude Micro Drone with MFCC and CNN,” Journal of the Korea Institute of Information and Communication Engineering, vol 24, no 3, pp. 364-370, 2020.

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

[9] YungHan Lee, “Speech/Audio Processing based on Deep Learning,” Broadcasting and Media Magazine, vol. 22, no. 1, pp. 46-57, 2017.

[10] Jeong-Ho Shin, Jin-Young Lee, Seong-Weon Seo, Kyung-Ah Sohn, “Similarity analysis of fashion image with color, style, and pattern based on VGG-19 model and histogram,” The Korean Institute of Information Scientists and Engineers Conf., pp. 2101-2103, 2018.

[11] Jae-Ho Kim, Jang-Young Kim, “Comparative analysis of performance of BI-LSTM and GRU algorithm for predicting the number of Covid-19 confirmed cases,” Journal of the Korea Institute of Information and Communication Engineering, vol 26, no 2, pp. 187-192, 2022.

[12] Byeong-Uk Jeon, JiSoo Kang, Kyungyong Chung., “AutoML and CNN-based Soft-voting Ensemble Classification Model For Road Traffic Emerging Risk Detection,” Journal of Convergence for Information Technology, vol. 11, no. 7, pp. 14-20. 2021.

저자소개

한승우 (Seung-Woo Han)

상명대학교 정보보안공학과 재학
관심분야 : 보안시스템, 정보보호
이메일 : han6502@daum.net

한성훈 (Seong-Hun Han)

상명대학교 정보보안공학과 재학
관심분야 : 보안시스템, 정보보호
이메일 : wk268@naver.com

유성민 (Seong-Min You)

상명대학교 정보보안공학과 재학
관심분야 : 보안시스템, 정보보호
이메일 : ysmuei@naver.com

송동호 (Dong-Ho Song)

상명대학교 정보보안공학과 재학
관심분야 : 보안시스템, 정보보호
이메일 : kingdh61@naver.com

서창진 (Chang-Jin Seo)

부산대학교 멀티미디어 공학박사
경력 : 센서기술연구소 연구원, 상명대학교 정보보안공학과 교수
관심분야 : 인공지능, 보안시스템, 악성코드분석
이메일 : cjseo@smu.ac.kr