심층신경만 기반 음성 특징 분석 기술을 활용한 딥보이스 탐지시스템

이현준, 순현상, 조성훈, 서창희, 강지윤, 유홍석(교신저자)
경운대학교 소프트웨어학부
e-mail:{bmjjl2, tnsgustkd, mc213213, wldnjs3608, wldbs5165}@naver.com,
hsyoo@ikw.ac.kr

Deep voice Detection System Using Deep Neural Network-Based Voice Feature Analysis Technology

Hyeon-Jun Lee, Hyeon-Sang Soon, Seong-Hun Jo, Chang-Hui Seo, Ji-Yun Kang,
Hong-seok Yoo(Corresponding Author)
School of Software, Kyungwoon University

요 약

최근 인공지능(AI) 기술의 발전은 다양한 분야에 긍정적인 영향을 이끌어내면서도, 동시에 부정적인 영향도 동반하고 있다. 특히, AI 기술을 이용한 보이스피싱(Voice Phishing) 사례가 급증하여 이에 대한 대처 방안이 요구되고 있다. 본 논문에서는 딥러닝 기반 음성 합성 기술인 딥보이스(Deep Voice)를 활용하여 음성 사기를 예방할 수 있는 애플리케이션을 제안한다. 본 논문에서는 MFCC와 Mel-Spectrogram을 이용해 음성의 특징을 추출하고 추출한 데이터를 바탕으로 VGG-19, BiLSTM모델을 이용해 모델을 생성 후 소프트 보팅을 통해 최종 결과 값을 판별하였다. 또한 추가적으로 성별에 따른 분류 모델을 통해 정확도를 향상하였다. 이를 통해 사용자는 녹음된 데이터를 분석하여 해당 데이터가 일반 음성인지 딥보이스로 생성된 음성인지를 식별할 수 있으며, 최종적으로 보이스피싱을 효과적으로 예방할 수 있을 것으로 기대된다.

▸ Keyword : 보이스피싱(Voice Phishing), 딥보이스(Deep Voice), 음성 특징 추출 기법, CNN(Convolutional Neural Network), BiLSTM(Bidirectional Long Short-Term Memory)

I. Introduction

최근 몇 년간 인공지능(AI) 기술을 악용한 범죄가 증가하고 있다. 특히, 딥보이스 기술을 이용해 지인을 사칭한 보이스피싱 피해 사례가 많이 증가하고 있다. 발전된 AI 기술과 접목하여 실제 음성과 구분하기 어려운 정교하고 신뢰성 있는 목소리를 보이스피싱 범죄에 이용하기 때문이다[1]. 이에 따른 보이스피싱 범죄를 예방할 수 있는 방안이 현재 요구되고 있는 상황이다. 본 논문에서는 남성과 여성의 음성의 주파수, 파형의 모양, 발음의 패턴 등에 차이를 활용하여 판별 정확성을 높일 수 있는 딥보이스 음성 탐지 방법을 제시한다. 남여의 음성적 차이에 따른 분류를 통한 모델 학습 진행시 모델이 패턴을 더 정확히 인식하고 구분할 수 있도록 돕는다[2]. 따라서 남성과 여성의 음성을 분리하여 음성 특징 추출을 진행하고 추출 데이터를 바탕으로 성별에 따른 딥러닝 모델 학습을 진행한다. 이를 통해, 딥보이스 음성과 일반 음성을 보다 정확하게 판별할 수 있는 탐지 시스템을 구축하였다.

II. Preliminaries

음성 특징 추출 기술로는 Librosa에서 제공하는 MFCC와 Mel-Spectrogram을 사용하였다. MFCC는 음성 신호에서 추출할 수 있는 음성데이터를 특징벡터화 해주는 알고리즘이다. 입력된 소리 전체를 대상으로 하는 것이 아닌 일정 구간 식 나누어 구간에 대한 스펙트럼을 분석하여 특징을 수치데이터 형태로 추출하는 기법이다. Mel-Spectrogram은 소리의 파형을 인간이 들을 수 있는 범위로 줄이는 기술인 Mel scale로 변환 후 파형을 그래프 이미지로 나타내는 추출 기법이다. 딥러닝 학습 모델은 Tensorflow의 keras를 이용해 VGG-19모델과 BiLSTM모델을 사용하였다. VGG-19모델은 컴퓨터 비전 분야에서 널리 사용되는 CNN 아키텍처로 총 19개의 레이어로 구성되어 있다. 다양한 시각적 계층을 통해 이미지 학습

에 강점이 있다. BiLSTM은 입력 시퀀스를 양방향으로 처리가 가능해 음성 인식과 같은 시계열 데이터 학습에 강점이 있다. 최종판별은 소프트 보팅 기법을 적용하였다. 소프트 보팅은 앙상블 기법중 하나로 각 판별 확률을 모두 더해 평균을 내어 최종 결과 값을 선정하는 방법이다.

III. Design and Development

1. System Architecture

Fig. 1. System 흐름을 나타낸다. 성별에 따른 딥러닝 모델을 생성하기 위해 남여 각각 일반 음성과 딥보이스 음성의 음성 특징을 추출한다. MFCC를 이용해 음성 특징 추출 시 음성 특징 정보를 갖는 수치 데이터 값들과 라벨링 정보를 가진 CSV파일이 생성된다. Mel-Spectrogram의 경우 노이즈 제거를 위해 흑백 처리 작업 후 음성 특징 정보를 가진 PNG 형식의 그래프 이미지 파일이 생성된다. 그 후 생성된 학습 데이터를 바탕으로 모델 학습을 진행한다. 음성의 특징 수치 데이터정보를 갖고 있는 CSV파일을 이용하여 BiLSTM모델의 학습이 진행되고 그래프 이미지 데이터는 VGG-19모델로 학습이 진행된다. 그 후 남여 각각 BiLSTM모델과 VGG-19모델이 하나씩 생성되고 생성된 모델을 통해 판별 음성과 성별 정보를 지정해주면 성별에 정보와 맞는 모델을 통해 음성이 딥보이스인지 아닌지 판별한다. 그 후 두 모델을 통한 음성에 대한 판별 확률을 소프트 보팅을 통해 최종판별 결과가 결정된다.

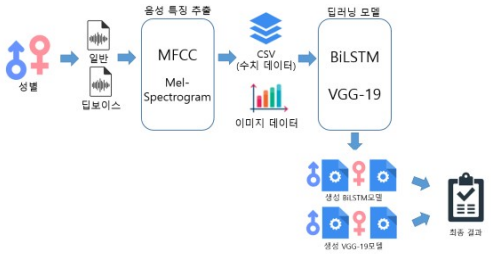

Fig. 1. Application Results

2. Deep learing model

학습에는 Ai-hub에서 제공하는 다화자 음성을 이용해 남여 각각 일반 음성 1000개 딥보이스 음성 1000개를 생성, 확보하여 진행하였다. 남여각각 딥보이스 음성과 일반 음성의 음성 특징을 추출하였다. Mel-Spectrogram은 그래프 이미지로 음성 주파수 정보를 인간의 청각 특성을 표현한 것이다. MFCC는 음성의 주파수 특징을 수치화하여 나타내는 알고리즘이다. MFCC와 Mel-Spectrogram모두 1초에 16000개의 샘플링 주파수를 추출하게 설정하였다. 그리고 윈도우 크기는 400, 분석 윈도우의 간격을 160, 추출 특징 개수 100개로 설정하여 0.01초마다 추출된 샘플링 주파수중 400개를 이용하여 100개의 음성 특징을 추출하게 하였다. 추가적으로 Mel-Spectrogram은 더 넓은 범위의 음량을 효과적으로 다를 수 있도록 데시벨로 변환을 진행한다. 최댓값을 0db로 설정하여 그래프 이미지에 표시될 때 다른 값들은 이에 상대적인 크기로 변환되어 표시되게 구현하였다. 이러한 과정을 거쳐 나온 음성 특징 데이터들을 바탕으로 학습을 진행하였다. 모델 학습에 있어서 그래프 이미지의 경우 이미지 학습에 강점이 있는 VGG-19 모델을 사용하여 학습 진행하였다. 뉴런의 개수는 512개, 과적합 방지를 위하여 Dropout를 0.3으로 학습 진행하였다. BiLSTM모델의 경우 뉴런의 개수는 512개, Dropout은 0.5로 학습 진행하였다. 그리고 두 모델 모두 출력층의 뉴런을 2로 설정하였고 학습과 검증데이터 비율을 8대2로 나누어 20번의 반복학습을 진행하였다. ModelCheckpoint를 이용해 반복학습 중 정확도 검증 시 정확도가 높은 모델을 h5파일로 저장하여 딥러닝 모델을 생성하였다.

3. Gender classification

성별 분류에 따른 모델 정확도 성능 향상 검사를 진행하였다. 남여 각각 직접 녹음한 일반 음성 100개와 새로 생성한 딥보이스 음성 100개로 검사 진행하였다. 남여 함께 학습한 모델의 경우 일반 음성과 딥보이스 음성 5대5 비율로 정확도 검사 진행하였다. 남여 음성을 함께 학습한 모델의 경우 정확도 95%의 정확도를 보였다. 그 후 남여 분리하여 학습을 진행한 모델의 정확도 검사의 경우 남자모델의 정확도는 98%, 여자모델의 경우 97.5%로 정확도 향상을 보였다.

Table 1. Improved Accuracy Performance

System	Male + Female	Male	Female
Accuracy	95%	98%	97.5%

IV. Conclusions

본 논문에서는 음성 특징 추출 알고리즘과 신경망 학습 모델을 성별에 따라 분류하여 학습을 진행해 정확도를 높인 딥보이스 음성 탐지시스템을 구현하였다. 이를 바탕으로 사용자 친화적으로 딥보이스를 보이스피싱 범죄 예방 애플리케이션을 구현하였다. 이러한 애플리케이션을 제공해 줌으로써 사용자가 딥보이스 음성 범죄를 구별하여 딥보이스 범죄 피해 사례가 줄어들 것이라 기대한다. 또한 군용, 금융도 시스템을 도입 가능하다 생각하고 충분한 학습 데이터양을 확보한다면 나이, 지역 등 더 다양한 분류 모델을 통해 더 정확도 높은 딥보이스 탐지시스템으로 발전 가능하다 생각한다.

Reference

[1] Lee hyuck kee “"brother! “Please save me” Your voice is being used for crime: AI voice phishing trap“, Thescoop, 2024

[2] Youn-ho Cho, & Kyu-sik Park (2008). A Study on The Improvement of Emotion Recognition by Gender Discrimination. The Institute of Electronics Engineers of Korea - Signal Processing, 45(4), 107-114.

[3] A Study on the Development of Deep Learning-Based Deep Voice Detection System Using Mel-Spectrogram and MFCC, http://journal.auric.kr/kieep/XmlViewer/f424187