딥러닝 기반 한국어 딥보이스 탐지 시스템

Korean Deep Voice Detection System Based on Deep Learning

정 우 진, 이 병 엽, 강 아 름
배재대학교

Jung Woo Jin, Lee Byoung-Yup, Kang Ah Reum
Pai Chai University

요약

딥보이스 기술이 발전해 감에 따라 딥보이스 음성과 실제 음성을 구별하기가 점차 어려워지고 있다. 과거부터 보이스피싱 피해 사례는 꾸준히 발생하고 있었지만, 최근에는 딥보이스 기술을 악용한 보이스피싱 피해 사례가 발생하고 있다. 하지만 이에 대한 대응책은 미흡한 실정이다. 본 논문에서는 딥러닝 기술을 활용한 한국어 딥보이스 탐지 시스템을 제안한다.

Ⅰ. Introduction

보이스피싱(voice phishing)이란 전화를 통해 기관이나 지인 등을 사칭하여 허위 사실을 말하면서 불안감을 조성해 송금을 요구하거나 개인정보를 수집하는 사기 수법을 의미한다. 2006년 5월 최초로 피해 사례가 집계된 이래, 현재 우리의 생활 속에서 경험할 수 있는 가장 흔한 범죄유형으로 자리 잡았다 [1]. 2023년 보이스피싱 피해액은 1,965억 원으로 전년 피해액인 1,451억 원 대비 약 35% 증가세를 보였다 [2]. 금융기관 및 수사기관에서는 매년 보이스피싱 피해를 줄이려고 각종 대응책을 내놓고 있지만, 보이스피싱 범죄의 수법과 기술이 나날이 진화해 감에 따라 대응책이 신유형의 보이스피싱을 따라가지 못하고 있는 것이 현상이다.

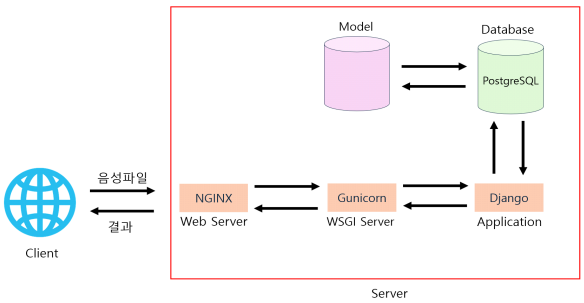
또한 최근에는 딥보이스 기술을 활용하여 특정인의 목소리나 말투를 학습시킨 후, 해당 인물을 사칭하여 금전적 피해를 발생시킨 피해 사례가 발생하고 있다. 이처럼 딥보이스 기술의 발전 속도에 맞춰 탐지 기술 또한 발전해야 하지만, 이와 관련된 선행 연구는 많지 않다.

본 논문에서는 한국어 딥보이스 음성을 탐지할 수 있는 시스템을 제안한다.

Ⅱ. Proposed System

1. Deep Voice Detection System Overview

본 논문에서 제안하는 시스템의 구성은 그림 1과 같다. Django Web Framework을 통해 웹 사이트를 구성한다. Django Web Framework이란 Python으로 작성된 오픈소스 웹 프레임워크(web framework)로 Python으로 작성된 모델 및 라이브러리와의 통합 및 연동이 용이하다.



▶▶ 그림 1. 딥보이스 탐지 시스템

브라우저의 정적 페이지 요청을 처리하고 동적 페이지 요청의 경우 WSGI 서버(web server gateway interface)를 호출하는 기능을 수행하는 웹 서버(web server)는 NGINX를 통해 구축한다. 웹 서버에서 Django 애플리케이션을 호출하는 기능을 수행하는 WSGI 서버는 Gunicorn을 사용한다. 데이터베이스(database)는 관계형 데이터베이스(SQL)의 일종인 PostgreSQL을 사용하여 구축하며, 사용자가 업로드한 음성 파일의 해쉬값(hash value)과 딥보이스 확률을 함께 저장하여 추후 동일한 음성파일의 딥보이스 확률 요청 시에 저장된 확률을 응답으로 제공해 서버의 자원 낭비를 방지한다.

궁극적으로 사용자에 의해 업로드된 음성파일은 앞선 모든 과정을 거쳐 마지막 단인 딥보이스 탐지 모델을 거쳐 확률값으로 반환한다.

2. Deep Voice Detection Model

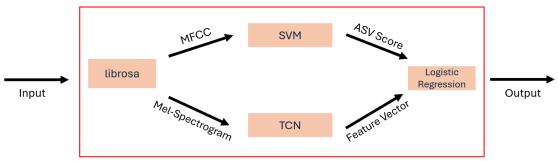
딥러닝 및 머신러닝에서는 원본 음성 데이터를 바로 학습에 사용할 수 없으므로 음성 특징 추출을 위해 Python 라이

브러리의 일종인 libsora를 사용하여 MFCC, Mel-Spectrogram을 추출한다 [3].

이때, MFCC(mel-frequency cepstral coefficients)란 음성 데이터에서 추출된 주파수의 특징을 수치화하여 나타낸 계수를 의미한다. Mel-Spectrogram이란 음성 데이터의 파형을 인간이 들을 수 있는 범위로 다운 스케일링한 후, 해당 파형을 그림으로 나타낸 모형을 의미한다. 본 논문에서 제안하는 딥보이스 탐지 모델은 그림 2와 같다.

MFCC는 머신러닝 기법 중 우수한 성능을 보이는 SVM(support vector machine) 모델을 거쳐 화자 검증(automatic speaker verification) 점수로 반환한다 [4].

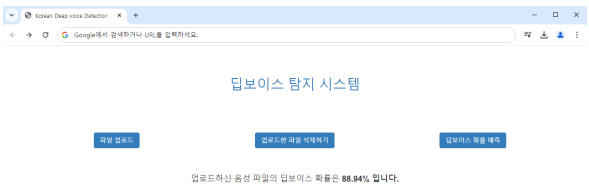
Mel-Spectrogram은 CNN의 일종인 TCN(temporal convolutional network) 모델을 사용하여 특징 벡터(feature vector) 형태로 반환한다. 앞선 단계에서 산출된 ASV Score와 Feature Vector를 확률모델의 일종인 로지스틱 회귀(logistic regression) 모델을 사용하여 확률 형태로 변환한다.



▶▶ 그림 2. 딥보이스 탐지 모델

3. Web Interface

사용자는 웹 사이트에 접속하여 딥보이스가 의심되는 음성파일을 서버로 전송한다. 서버 단에서는 딥보이스 탐지 모델을 통해 전송된 음성파일의 딥보이스 확률을 시각적으로 출력한다. 그림 3은 사용자가 음성 파일을 올린 후, 딥보이스 확률이 화면에 출력되는 장면이다.



▶▶ 그림 3. 음성파일 업로드 후 결과를 출력하는 화면

Ⅲ. Conclusion

본 논문은 한국어 딥보이스 탐지 시스템을 제안한다. MFCC를 머신러닝 기법인 SVM을 통과시켜 ASV Score로 변환하고, Mel-Spectrogram은 딥러닝 기법인 TCN을 사용하여 Feature Vector 형태로 변환한다. 그 결과를 바탕으로 Logistic Regression을 수행하여 음성파일의 딥보이스 확률을 예측할 수 있다.

본 논문에서 제안한 한국어 딥보이스 탐지 시스템은 보이스피싱 탐지, 딥보이스를 악용한 가짜뉴스 탐지 등 다양한 분야에서 활용될 수 있으며, 별도의 애플리케이션 설치를 필요로 하지 않는 웹 사이트를 기반으로 구성되어 모바일, PC 등 다양한 플랫폼(platform)에서 쉽게 사용할 수 있다.

향후 연구에서는 한국어 대화 데이터 셋과 한국어 딥보이스 데이터 셋을 수집하고 본 시스템을 구현하여 실제 서비스화에 이를 수 있을 것으로 예상된다. 이로써 사회적인 문제로 제기되는 딥보이스 관련 범죄 피해를 감소시키는 데 크게 이바지할 것으로 기대된다.

Acknowledgment

Following are results of a study on the “Convergence and Open Sharing System” Project, supported by the Ministry of Education and National Research Foundation of Korea.

References

[1] Cho, H.-D. (2012, July 28). “Voice Phishing Occurrence and Counterplan”. The Journal of the Korea Contents Association. The Korea Contents Association. https://doi.org/10.5392/jkca.2012.12.07.176

[2] 금융감독원 (2024.3.8.). “2023년 보이스피싱 피해현황 분석”. 보도자료.

[3] 한승우, 한성훈, 유성민, 송동호, 서창진. (2023). “Mel-Spectrogram과 MFCC를 이용한 딥러닝 기반 딥보이스 탐지시스템 개발에 관한 연구”. 전기학회논문지 P, 72P(3), 186-192, 10.5370/KIEEP.2023.72.3.186

[4] S. Raghavan, G. Lazarou and J. Picone, “Speaker Verification using Support Vector Machines”, Proceedings of the IEEE SoutheastCon 2006, Memphis, TN, USA, 2006, pp. 188-191, doi: 10.1109/second.2006.1629347.