# KoBERT를 활용한 실시간 보이스피싱 탐지기법 개념설계

김영진[O], 이병엽[*], 강아름[*]
[*]배재대학교 사이버보안학과,
[O]배재대학교 스마트ICT융합학과
e-mail: jin971001@naver.com[O], {bylee, armk}@pcu.ac.kr[*]

# Design of Real-Time Voice Phishing Detection Techniques using KoBERT

Yeong Jin Kim[O], Byoung-Yup Lee[*], Ah Reum Kang[*]
[*]Dept. of Cyber Security, Pai Chai University,
[O]Dept. of Smart ICT Convergence, Pai Chai University

● 요 약 ●

본 논문은 금융 범죄 중 하나인 보이스피싱을 실시간으로 예방하기 위한 탐지 기법을 제안한다. 제안된 모델은 수화기에 출력되는 음성을 녹음하고 네이버 CSR(Cloud Speech Recognition)을 통해 텍스트 파일로 변환한 후 딥러닝 기반의 KoBERT를 바탕으로 다양한 보이스피싱 패턴을 학습하여 실시간 환경에서의 신속하고 정확한 탐지를 위해 실제 통화 데이터를 적절하게 처리하여, 이를 통해 효과적인 보이스피싱 예방에 도움을 줄 것으로 예상된다.

**키워드:** 보이스피싱(Voice Phishing), 네이버 CSR(Cloud Speech Recognition), KoBERT, 실시간 탐지

## I. Introduction

보이스피싱(Voice Phishing)이란 피싱(Phishing)과 보이스(Voice)의 결합으로 이루어진 단어이다. 피싱의 어원은 낚시에서 유래하였는데, 공격자가 가짜로 만든 미끼를 사용하여 피해자를 끌어들이듯이 민감한 정보를 얻으려는 행위를 비유적으로 나타낸 용어이다. 보이스피싱은 피싱이라는 단어 앞에 음성을 의미하는 보이스가 붙은 용어이다. 보이스피싱은 전화 통화를 통한 공격 형태로 휴대전화 등 전기 통신수단을 이용하여 타인을 속이거나 협박함으로써 자금을 송금, 이체하도록 하게 하거나 개인정보를 알아내어 자금을 송금, 이체하는 행위이다[1].

공공데이터 포털에서 구할 수 있는 경찰청 보이스피싱 현황 자료에 따르면 연간 보이스피싱 발생 건수는 2019년 37,677건, 2020년 31,681건, 2021년 30,982건으로 연간 3만 건 이상의 보이스피싱 범죄가 꾸준히 발생하고 있으며, 피해 금액은 Table 1과 같이 2019년 6,398억 원, 2020년 7,000억 원, 2021년 7,744억 원으로 피해 금액이 점점 늘어나고 있다[2]. 보이스피싱은 주로 전화 통화를 통해 이루어지는 만큼 이를 실시간으로 탐지하고 차단하는 것이 보이스피싱 예방에 핵심이라고 할 수 있다.

Table 1. Voice Phishing Status

| Year | Occurrences | Financial Loss (KRW) |
|---|---|---|
| 2019 | 37677 | 639.8 billion |
| 2020 | 31681 | 700 billion |
| 2021 | 30982 | 774.4 billion |
| 2022 | 21832 | 543.8 billion |

이에 본 논문에서는 기존의 보이스피싱 탐지 시스템에 대해 살펴보고, 새로운 실시간 보이스피싱 탐지 시스템에 대한 설계를 제안한다.

## II. Related Work

현재 보이스피싱 실시간 탐지 시스템을 찾아보면 악성 앱 탐지, 문자 의심 키워드 발췌, 이상 URL 탐지 혹은 ATM 거래 이상행동 위주의 탐지가 가장 많이 존재한다.

신한은행은 고객이 ATM 거래 중 휴대전화 통화를 하거나 선글라스 및 모자를 착용하는 등의 보이스피싱 데이터를 분석해 얻은 유사 행동을 보일 경우 이를 탐지해 주의 문구를 안내하는 시스템을 국내 최초로 도입하였다[3].

구승연 외는 MFCC(Mel-Frequency Cepstral Coefficient)를 통해 음성 신호를 전처리하고 LSTM(Long Short-Term Memory)을 이용한 보이스피싱 판별 알고리즘을 설계하였다[4]. 해당 알고리즘의 음성 파일 MFCC 테스트 정확도는 96.22%, 실시간 음성에 대한 정확도는 96.70%로 나타났다.

## III. System Overview

본 논문에서 제안하는 보이스피싱 실시간 탐지 시스템을 Fig. 1과 같이 구성하였다.

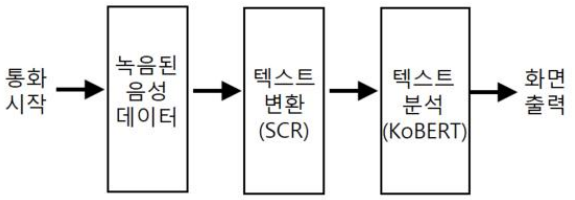


Fig. 1. System Overview

해당 보이스피싱 실시간 탐지 시스템은 사용자의 휴대전화에 설치되어 실행하는 애플리케이션이며 동작 과정은 다음과 같다.

첫 번째, 사용자가 통화를 시작하면 자동으로 통화 음성을 녹음하여 서버에 전송한다. 음성은 최대 20초 길이로 나누어서 녹음되는데, 이는 텍스트로 변환하는 작업을 진행하는 API인 네이버 Cloud Platform에 있는 CSR(Cloud Speech Recognition)에서 지원하는 파일의 길이가 최대 20초이기 때문이다.

두 번째, 20초 길이로 나누어서 녹음된 음성 파일은 네이버 CSR을 이용하여 텍스트 변환을 진행한다.

세 번째, 변환된 텍스트 데이터를 분석하여 해당 통화 내용이 보이스피싱인지 판별한다. 판별 과정에는 KoBERT를 사용한다.

KoBERT는 기존 BERT의 한국어 성능 한계를 극복하기 위해 SKT Brain에서 개발하였다. 한국어 위키에서 500만 개의 문장과 5,400만 개의 단어를 한국어의 불규칙한 언어 변화의 특성을 반영하기 위한 데이터 기반 토큰화 기법을 적용하였다. 주로 감정 분류 모델과 감성 분석에 많이 쓰이는 자연어 처리 모델이다[5].

마지막으로 KoBERT를 통해 보이스피싱인지 아닌지 판별하고, 보이스피싱으로 판별되면 사용자의 화면에 경고 메시지를 출력한다.

## IV. Conclusions

본 논문에서는 보이스피싱의 실시간 탐지를 위해 네이버 CSR API와 KoBERT가 결합한 탐지 시스템을 제안한다. 이 실시간 보이스피싱 탐지 시스템은 기존의 유사 연구에서 확장한 개념으로 한글 STT 변환율이 가장 높은 네이버 CSR과 가장 높은 정확도를 보이는 한글 자연어 처리 모델인 KoBERT를 결합하여 기존의 연구보다 빠르고 정확도가 높은 결과를 제공할 것으로 예상된다. 추후 본 설계안을 바탕으로 실제로 구현되어 동작하는 실시간 보이스피싱 탐지 시스템을 개발해 보고자 한다.

## ACKNOWLEDGEMENT

Following are result 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] KOREA FEDERATION OF BANKS, *https://portal.kfb.or.kr/voice/vphishing.php*

[2] Data Portal, *https://www.data.go.kr/data/15063815/fileData.do*

[3] FORTUNE KOREA, *http://www.fortunekorea.co.kr/news/articleView.html?idxno=21586*

[4] Seung-Yeon Koo, You-Jin Lee, Jung-Hwa Kang, Jae-Hyun Kim, “Design of Phone Scam Discrimination Algorithm using LSTM.”Proceedings of Symposium of the Korean Institute of communications and Information Sciences, pp. 989-990. 2021.

[5] SKTBrain KoBERT, *https://github.com/SKTBrain/KoBERT*