Journal of KIIT. Vol. 23, No. 10, pp. 175-183, Oct. 31, 2025. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2025.23.10.175

KoBERT와 CNN-LSTM을 활용한 보이스피싱 탐지 시스템

김지원*[1], 구슬이*[2], 김혜진**, 강창구***

Voice Phishing Detection using KoBERT and CNN-LSTM

Jiwon Kim*[1], Seuli Gu*[2], Hyejin Kim**, and Changgu Kang***

요 약

보이스피싱은 수법의 지능화와 정교화로 인해 심각한 사회적 문제를 야기하고 있으며, 이를 효과적으로 탐지하기 위해서는 통화 내용의 문맥적 위험도와 발화 음성의 위변조 여부를 종합적으로 판단하는 기술이 필수적이다. 기존의 키워드 기반 탐지 방식은 새로운 피싱 시나리오에 취약하고, 단일 모델만으로는 복합적인 공격을 방어하는 데 한계가 있다. 본 논문에서는 텍스트와 음성을 병렬적으로 분석하는 듀얼 파이프라인 구조의 보이스피싱 탐지 시스템을 제안한다. 텍스트 분석에는 KoBERT 기반을 통해 문맥적 위험도를 산출하며, 음성 분석에는 CNN-LSTM 하이브리드 모델을 활용하여 합성 음성 여부를 판별한다. 제안된 시스템은 금융감독원의 실제 피싱 사례 데이터와 so-vits-svc-fork 기반의 합성 음성을 포함한 학습 데이터를 바탕으로 훈련되었으며, 실제 피싱 샘플에 대해서는 94.9%의 높은 탐지율을 기록하였다.

Abstract

Voice phishing has emerged as a serious social issue due to increasingly intelligent and sophisticated tactics. To effectively detect such threats, it is essential to assess both the contextual risk of conversations and the authenticity of the speaker's voice. This paper proposes a dual-pipeline detection system that simultaneously analyzes text and audio. The text pipeline utilizes KoBERT to estimate contextual risk, while the audio pipeline employs a CNN-LSTM hybrid model to detect synthetic speech. Trained on real-world phishing cases from the Financial Supervisory Service and synthetic data generated with the so-vits-svc-fork framework, the proposed system achieved a detection accuracy of 94.9%.

Keywords
voice phishing, KoBERT, voice forgery, deep learning, text analysis

* 경상국립대학교 컴퓨터공학부 학사과정
- ORCID[1]: https://orcid.org/0009-0009-5266-5008
- ORCID[2]: https://orcid.org/0009-0007-1784-8971
** 서울연구원 연구위원
- ORCID: https://orcid.org/0000-0001-6541-3701
*** 경상국립대학교 컴퓨터공학부 교수(교신저자)
- ORCID: https://orcid.org/0000-0003-4060-6835

· Received: Jul. 28, 2025, Revised: Aug. 07, 2025, Accepted: Aug. 10, 2025
· Corresponding Author: Changgu Kang
Dept. of Computer Science and Engineering, Gyeongsang National University, Korea
Tel.: +82-55-772-3321, Email: cgk@gnu.ac.kr

Ⅰ. 서 론

보이스피싱은 전화나 음성 메시지를 통해 개인정보와 금융 정보를 탈취하는 대표적인 사회공학적 범죄로, 최근 합성 음성 기술의 발전과 함께 수법이 더욱 정교화되고 있다. 실제로 국내에서 보이스피싱 피해 사례는 매년 증가하고 있으며, 비대면 금융 환경의 확대와 맞물려 그 위협은 점차 심각해지고 있다[1].

보이스피싱은 단순한 기술적 해킹이 아니라, 인간의 인지 편향과 신뢰 메커니즘을 악용하는 사회공학적 범죄로 분류된다. 이러한 유형의 공격은 사용자의 심리적 취약점을 표적으로 삼기 때문에 기존의 기술 중심 보안 체계만으로는 효과적인 대응에 한계가 있다[2].

최근 합성 음성 기술과 대규모 언어 모델[3]의 발전으로, 보이스피싱 수법은 실제 대화와 구분이 어려울 정도로 정교해지고 있다. 특히 음성, 텍스트뿐만 아니라 이미지, 동영상, 딥페이크 등 다양한 인공지능 기술과 결합되면서 공격 양상은 더욱 다양화되고 있다[4]. 그럼에도 불구하고, 보이스피싱 피해의 대부분은 여전히 전화 통화나 음성 메시지 등 음성 기반 매체를 통해 발생하며, 실시간 상호작용 중 설득 과정에서 피해가 집중되고 있다[5].

이러한 문제를 해결하기 위해, 본 연구에서는 텍스트와 음성 정보를 결합한 보이스피싱 탐지 시스템을 제안한다. 제안된 시스템은 KoBERT 기반의 텍스트 분류기와 CNN-LSTMConvolutional Neural Network – Long Short-Term Memory) 기반의 합성 음성 탐지기로 구성되어 있으며, 입력된 음성은 Resemblyzer[6]를 활용한 화자 임베딩 및 K-means 클러스터링을 통해 화자를 분리한 후, Whisper STT[7]를 통해 텍스트로 변환된다. 이를 통해 기존 키워드 필터링이나 텍스트 분류 중심 기술의 한계를 극복하고, 실제 통화 기반의 실시간 탐지와 문맥 분석이 가능한 지능형 탐지 모델을 구현하고자 하였다.

합성 탐지를 위해 한국어 음성 데이터를 학습하여 so-vits-svc-fork[8]를 기반으로 음성을 합성하였으며, 실제 음성과의 분류 학습을 통해 탐지 모델을 훈련하였다. 또한 금융감독원에서 제공한 실제 보이스피싱 음성 데이터를 STT로 변환하여 텍스트 학습 데이터로 활용하고, AI Hub[9]의 한국인 대화 음성과 일상대화 한국어 멀티세션 데이터를 테스트셋으로 구성함으로써, 실제 환경을 반영한 탐지 성능을 평가하였다.

Ⅱ. 관련 연구

보이스피싱 탐지에 관한 기존 연구는 주로 텍스트 분류 기반 모델, 음성 기반 딥러닝 기법, 합성 음성 판별, 그리고 모바일 환경에서의 보안 체계에 중점을 두고 이루어졌다. 본 장에서는 주요 선행연구들을 소개하고, 본 연구가 이들과 비교해 어떤 기술적 차별성과 실용적 기여를 갖는지를 논의한다.

M. K. M. Boussougou and D.-J. Park[10]은 FastText 임베딩과 어텐션 기반 1D CNN-BiLSTM (Convolutional Neural Network – Bidirectional Long Short-Term Memory) 구조를 결합하여, 보이스피싱 문장의 구조적 특징을 학습하는 하이브리드 모델을 제안하였다. 이 모델은 문맥 연산과 시계열 학습을 동시에 수행함으로써 성능을 개선하였다. 그러나 두 연구 모두 실제 보이스피싱 통화로부터 직접 수집된 텍스트 데이터를 사용하지 않고, 정제된 스크립트 기반의 데이터셋에 의존하였다는 점에서 실제 환경 반영이 제한적이었다.

합성 음성 탐지에 대한 연구로는 B. Zhang et al.[11]이 발표한 리뷰 논문에서 CNN, ResNet, LSTM 계열 모델을 이용한 다양한 접근 방식의 정확도와 한계점을 분석하였다. 해당 논문은 음성 특징 추출(MFCC, Mel-Frequency Cepstral Coefficients), 상향 정규화된 켑스트럼 계수(CQCC, Constant Q Cepstral Coefficient), 스펙트로그램(Spectrogram) 등 음향 특징 기반 탐지 기법들이 여전히 합성 음성에 취약함을 지적하며, 고차원 시계열 정보를 반영한 신경망 구조의 필요성을 강조하였다. 또한 A. Triantafyllopoulos et al.[12]은 사회공학 기반 음성 사기의 심리적 설득 메커니즘과 사용자 반응 특성을 분석하였으며, 음성 내 억양(Ethnic accent) 등이 수신자의 신뢰 판단에 큰 영향을 미침을 실험적으로 제시하였다. 이는 본 연구에서 합성 음성을 실제 화자의 억양과 스타일을 반영하여 생성하고, 그 탐지

여부를 평가하는 구성과 같다.

J. Kim et al.[13]은 Android 기반 보이스피싱 악성 앱들의 동작을 분석하고, 시스템 레벨에서 이를 탐지 및 차단하는 HearMeOut 프레임워크를 제안하였다. 이들은 1,017개의 실 앱을 분석해 콜 리디렉션, 콜 화면 오버레이, 합성 음성 재생 등 다양한 기능이 보이스피싱 성공률을 높이는 데 기여함을 밝혔으며, 시스템 API 호출을 탐지하여 이러한 행동을 차단할 수 있음을 입증하였다.

본 연구는 기존 연구들의 한계를 보완하여, 다음과 같은 차별화된 요소를 통해 현실성 있는 보이스피싱 탐지 시스템을 제안한다. 본 연구는 기존의 단일모달 보이스피싱 탐지 기술의 한계를 극복하기 위해, 음성과 텍스트 정보를 동시에 활용하는 구조를 도입하였다. 구체적으로는 CNN-LSTM 기반의 합성 음성 탐지 모델과 KoBERT 기반의 텍스트 분류 모델을 병렬적으로 적용한 후, 두 분석 결과를 80:20의 비율로 가중 평균하여 최종 판단을 수행한다. 텍스트 분석에는 문장 내 단어 간의 의미적 관계를 반영할 수 있는 Self-Attention 기반 KoBERT 구조를 도입하였다. 기존 단어 중심 접근 방식은 단어의 단순 출현 여부에 의존하기 때문에, '계좌', '송금'과 같은 단어가 평범한 문맥에서 사용될 경우에도 높은 위험도를 부여하는 오류가 있었다. 반면 Self-Attention 메커니즘은 문장 내 단어 간의 문맥적 관계를 계산하여, 피싱 의도를 보다 정교하게 식별할 수 있는 장점을 제공한다. 텍스트 데이터는 금융감독원이 제공한 실제 보이스피싱 음성 자료를 Whisper STT 엔진을 통해 텍스트로 변환하여 확보하였으며, 음성 데이터는 AI Hub의 실제 한국어 대화 데이터를 기반으로 so-vits-svc-fork 모델을 활용해 합성 음성을 생성한 후, 이를 딥보이스 탐지 모델의 학습에 활용하였다. 또한, 통화 중 등장하는 다수의 화자를 효과적으로 분리하기 위해 Resemblyzer와 KMeans 클러스터링 기법을 적용하여 화자 임베딩을 생성하고, 각 화자의 발화를 개별 분석함으로써 다중화자 환경에서도 안정적인 탐지가 가능하도록 하였다.

III. 보이스피싱 탐지시스템

본 장에서는 제안하는 보이스피싱 탐지 시스템의 전체 구조와 각 구성 요소의 세부 동작 원리를 설명한다. 그림 1은 제안된 시스템의 전체 흐름을 제안한다. 입력된 통화 음성은 화자 분리를 거친 후, 각 화자 단위로 텍스트 분석과 음성 분석이 병렬적으로 수행되며, 이후 통합된 점수 기반으로 보이스피싱 여부를 최종 판단한다.

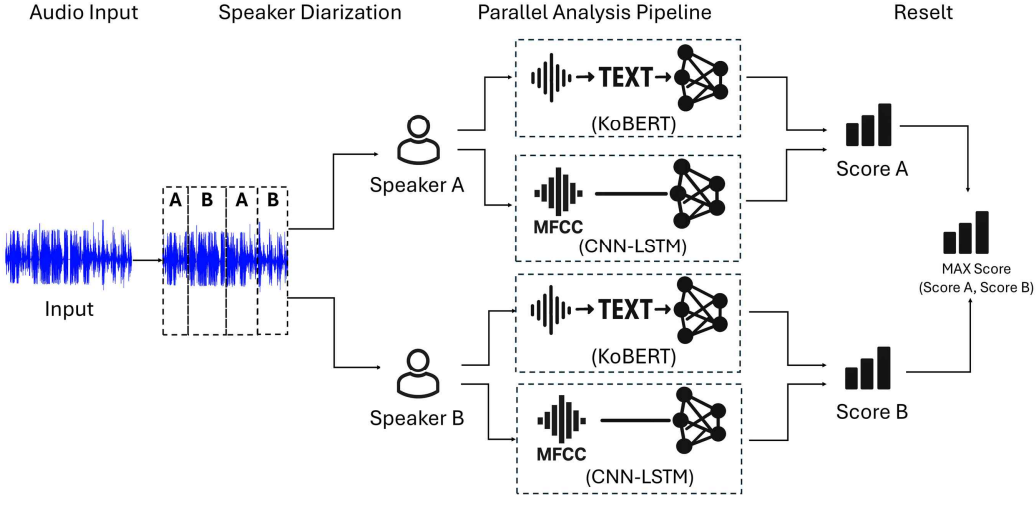

그림 1. 보이스피싱 탐지시스템의 전체 구조
Fig. 1. Architecture of the voice phishing detection system

3.1 화자 분리 및 전처리

본 논문의 화자 분리 모듈은 Resemblyzer를 활용하여 256차원 음성 임베딩 벡터를 추출한 뒤, K-평균 군집화 기법을 적용하여 화자를 분리하는 방식으로 구현하였다. 군집의 개수는 실제 통화 음성 데이터의 분포와 화자 분리의 신뢰도를 고려해 두 개로 고정하였다. K-mean 군집화 이후 각 군집에 포함된 음성 세그먼트의 개수가 15개 미만이거나, 전체 임베딩 샘플 수가 두 개 미만인 경우에는 화자 분리가 불안정하다고 판단하여 전체 오디오를 하나의 화자로 처리하였다. K-mean 알고리즘의 무작위 시드는 0으로, 초기화 반복 횟수는 자동으로 설정하였으며, 분리된 각 화자의 음성 구간은 각각 음성 텍스트 변환 단계와 딥보이스 판별 단계로 전달된다. 군집 수를 동적으로 결정하기 위한 추가적인 평가 지표는 적용하지 않았으며, 실제 현업 데이터를 대상으로 한 실험 결과, 이러한 규칙 기반의 분리 방식이 과도한 분할이나 오류를 방지하는 데 가장 효과적인 것으로 나타났다. 또한 임베딩 벡터의 분포가 명확하지 않거나 군집화 과정에서 오류가 발생하는 경우에도 전체 오디오를 하나의 화자로 처리하는 예외 처리를 적용하였다.

3.2 텍스트 변환 및 KoBERT 기반 텍스트 탐지

분리된 화자의 음성 구간은 OpenAI Whisper 음성 인식 엔진을 통해 문자 데이터로 변환된다. 변환된 텍스트는 특수문자 제거, 불용어 제거, 형태소 분석 및 단어 단위 분할 과정을 포함한 전처리를 거친다. 이후 정제된 문장은 사전 학습된 KoBERT 기반 분류 모델에 입력되어, 보이스피싱 여부를 0부터 100까지의 점수로 산출한다. 해당 모델은 768차원의 문장 표현 벡터를 생성한 뒤, 다층 연결 신경망과 확률 예측 함수를 거쳐 이진 분류 결과를 도출하며, 이를 텍스트 기반 위험 점수(Text score)라고 정의한다. 이때 Attention 연산은 식 (1)과 같이 정의된다.

$$Attention(Q,K,V) = softmax(\frac{QT^T}{\sqrt{d_k}})V \qquad (1)$$

그림 2는 이러한 KoBERT 기반 텍스트 분류 모델의 전체 구조를 제안한다. 입력 문장은 토크나이징과 임베딩 과정을 거친 후, 단어 간 관계를 학습하는 Self-Attention 기반 Transformer 인코더를 통해 문맥 정보를 학습하고, 이후 다층 연결 신경망을 통해 최종 분류 결과를 산출한다.

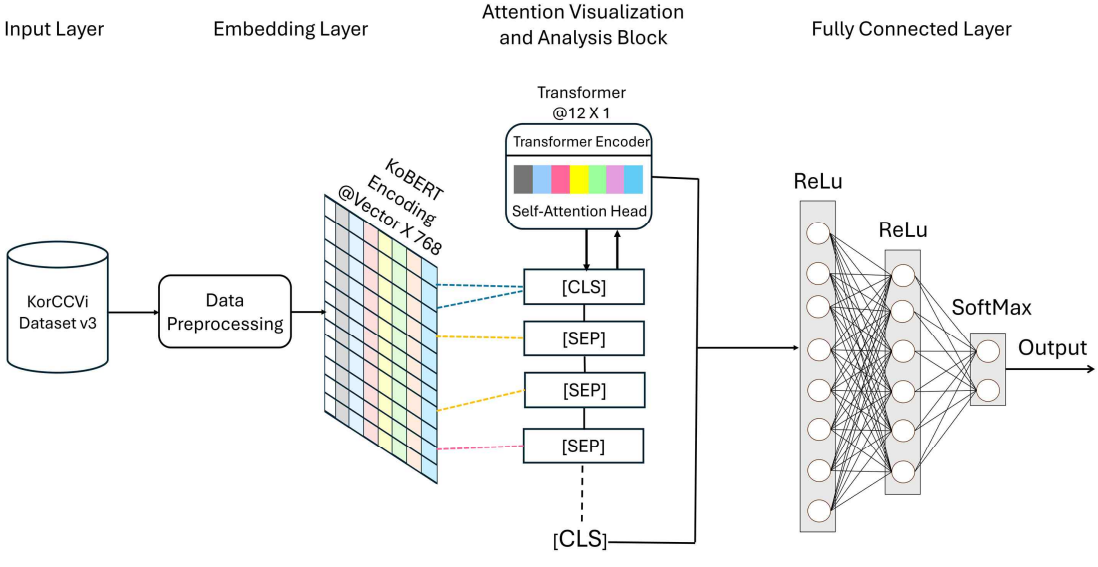

그림 2. KoBERT 기반 텍스트 분류 모델 구조
Fig. 2. Architecture of the KoBERT-Based text classification model

한국어 보이스피싱 텍스트는 전처리 후 KoBERT 임베딩을 거쳐 Self-Attention 기반 Transformer Encoder에서 문맥 정보를 학습하고, 최종적으로 Fully Connected Layer를 통해 피싱 여부를 이진 분류한다.

텍스트 분석 모델로 KoBERT의 Self-Attention 메커니즘을 선택한 이유는, 단어 간의 단순 빈도나 출현 여부만으로는 보이스피싱을 판별하기 어려운 구조적 특성 때문이다. 예를 들어 '계좌'와 같은 단어는 비피싱 문맥에서도 자주 등장하지만, '검찰청', '안전', '이체'와 같은 단어와의 동시 출현 여부에 따라 위험도가 급격히 상승한다. Self-Attention은 문장 내 모든 단어의 상호 관계를 고려하여 문맥에 따른 의미 변화를 반영하므로, 문장 내 단어 배열이 복잡하거나 의도적으로 조작된 문장을 탐지하는 데 강력한 성능을 발휘한다. 반면 기존 RNN은 긴 문장에서 초반 정보의 영향력이 감소하며, CNN은 멀리 떨어진 단어 사이의 의미 연계를 처리하기 어려워 본 연구에서는 Self-Attention 기반 구조를 제안하였다.

3.3 CNN-LSTM 기반 합성 음성 탐지

동일한 화자의 음성 구간은 합성 여부를 판별하기 위해 음성 특성 분석 모델에 입력된다. 본 연구에서는 시간 및 주파수 기반 특징을 효과적으로 분석할 수 있도록 합성곱 신경망(CNN)과 순환 신경망(LSTM)을 결합한 혼합 구조를 제안한다.

먼저, 입력된 음성 구간으로부터 음성 특징 추출을 추출하였으며, 이때 분석 창 크기는 25밀리초, 이동 간격은 10밀리초로 설정하였다. 음성 특징 추출은 사람의 청각 특성을 반영한 대표적인 음향 특징으로, 음성의 주파수 분포와 화자의 발화 패턴을 효과적으로 나타낼 수 있다. 이렇게 추출된 음성 특징 추출 시퀀스는 먼저 일차원 합성곱 신경망 계층에 입력되어, 짧은 시간 구간 내에 존재하는 미세한 음향적 변화를 학습한다. 합성곱 신경망은 지역적인 특징을 포착하는 데 강점을 가지며, 음성 파형 내의 잡음, 왜곡, 높낮이 변화 등 합성 음성에서 흔히 나타나는 이상 패턴을 감지하는 데 효과적이다. 이러한 지역적 특징 추출 과정은 식 (2)와 같이 정의된다.

$$y(t) = \sum_{i=0}^{k-1} x(t+i) \cdot w(i) \tag{2}$$

양방향 장기 기억 순환 신경망(BiLSTM) 계층은 시간의 흐름을 따라 앞뒤 문맥 정보를 동시에 학습하여, 발화의 억양, 리듬, 구조 등 시계열 패턴을 효과적으로 반영한다. 여기에 어텐션 메커니즘을 도입하여, 발화 중 중요한 구간에 가중치를 부여하고 모델이 핵심 프레임에 집중할 수 있도록 설계하였다.

이 과정을 통해 출력되는 값은 해당 음성이 합성 음성일 확률을 의미하는 음성 기반 위험 점수로 정의된다. 그림 3은 CNN-LSTM 기반 합성 판별기의 전체 구조를 시각적으로 제시하며, 음성 특징 추출부터 합성곱 신경망, 순환 신경망, 어텐션, 출력 계층으로 이어지는 흐름을 통해 시간적·공간적 특징을 동시에 반영한다.

모델 학습에는 AI Hub의 실제 한국어 발화 데이터를 '실제 음성'으로, 동일 데이터를 기반으로 so-vits-svc-fork 모델을 활용해 생성한 음성을 '합성 음성'으로 정의하여 총 약 18,000개의 샘플로 이진 분류 학습을 수행하였다(학습:검증 = 8:2).

합성 음성 데이터의 품질 검증을 위해, 실제-합성 1:1 쌍에 대해 MFCC 평균 벡터 간 코사인 유사도를 산출하였으며, 1,691쌍 기준 평균 유사도는 0.974로 매우 높게 나타났다. 이는 합성 음성이 실제 음성과 스펙트럼 및 음색 특성 측면에서 매우 유사함을 객관적으로 확인한 결과이다. 또한, 실험적으로 합성 음성의 청취 평가(MOS)도 실시하였으며, 실제 음성과 합성 음성 모두에서 청취자가 높은 자연스러움을 보고하였다.

이러한 검증 과정을 통해 확보된 데이터는 실제-합성 간 미세한 차이까지 정교하게 학습할 수 있는 기반을 제공하였으며, 모델의 판별 신뢰도 향상에 중요한 역할을 하였다.

3.4 점수 통합 및 최종 판단

제안된 보이스피싱 탐지 시스템은 텍스트 분석 결과와 음성 분석 결과를 통합하여 최종 판단을 수행한다. 각각의 분석 결과는 0부터 100 사이의 점수

로 산출되며, 텍스트 분석에서 도출된 점수는 텍스트 기반 위험 점수(Text score), 음성 분석에서 도출된 점수는 음성 기반 위험 점수(Voice score)로 정의된다. 이 두 점수는 가중 평균 방식으로 통합되며, 최종 점수(Final score)는 식 (3)과 같이 계산된다.

$$FinalScore = 0.8 \cdot TextScore + 0.2 \cdot VoiceScore \quad (3)$$

3.5 시스템 구성 요약

제안된 보이스피싱 탐지 시스템은 음성 입력부터 최종 판단까지 전 과정을 자동화한 통합 구조로 구성되어 있다. 화자 분리 모듈을 통해 다중 화자의 음성을 분리한 뒤, 각 화자 단위로 텍스트 분석과 음성 분석을 병렬적으로 수행한다. 산출된 위험 점수는 텍스트:음성 = 8:2의 가중치로 통합되며, 모든 화자 중 가장 높은 점수를 기준으로 최종 판단을 내리도록 설계되어 있다. 시스템은 기능별로 모듈화되어 있어 유지보수와 기능 개선이 용이하다.

특히 본 시스템은 사전 정의된 시나리오가 아닌, 실제 환경에서 수집된 음성 데이터를 기반으로 구성되었다. 텍스트 분석에는 금융감독원 보이스피싱 사례 음성을 Whisper를 통해 변환한 텍스트를 활용하였고, 음성 분석에는 AI Hub의 한국어 대화 데이터를 바탕으로 딥보이스 합성을 수행하여 학습 데이터를 구성하였다. 이러한 실환경 기반의 설계는 시스템의 적용 가능성을 높이며, 실제 금융 사기 방지에 활용 가능한 실효성 있는 탐지 솔루션으로 발전할 수 있는 가능성을 제시한다.

IV. 실 험

4.1 실험 데이터 구성 및 절차

제안하는 보이스피싱 탐지 시스템의 성능을 평가하기 위해 실제 환경 기반의 테스트 데이터를 구축하였다. 텍스트 분석 모델의 경우, 금융감독원에서 제공하는 보이스피싱 사례 음성을 Whisper STT를 통해 변환하여 피싱 텍스트 데이터를 생성하였고, 일반 대화 데이터는 AI Hub의 한국어 일상 대화 데이터를 활용하였다. 전체 25,000개의 문장을 수집하여 학습과 검증에 사용하였으며, 실제 테스트셋은 총 200개의 실통화 기반 음성 데이터를 사용하여 평가를 수행하였다.

음성 기반 딥보이스 탐지 모델은 AI Hub의 실제 화자 음성을 바탕으로 학습된 so-vits-svc-fork 모델을 통해 합성 음성을 생성하여, 실제-합성 쌍으로 구성된 총 18,000개의 오디오 샘플을 학습에 활용하였다. 학습:검증 비율은 8:2로 구성하였으며, 모델은 10 에포크(epoch) 학습 후 수렴하였다.

4.2 실험 결과 및 분석

보이스 피싱 탐지 시스템의 최종 성능은 다음과 같다. 실제 보이스피싱 음성 39개 중 37개를 정확히 탐지하여 재현율 94.9%를 기록하였고, 정상 음성 61개 중 47개를 정상으로 판단하여 전체 정확도는 84.0%에 도달하였다. 일부 정상 음성이 보이스피싱으로 과탐지된 사례도 존재했으나, 이는 본 시스템이 보수적인 판단 기준(최종 점수 90 이상: 확정)을 적용한 결과이다. 본 연구의 목표가 피싱 음성을 놓치지 않고 탐지하는 데 있음에 따라, 높은 재현율 확보가 실질적 응용 측면에서 더 중요한 판단 요소가 된다.

그림 3은 위 테스트 결과를 기반으로 생성된 혼동행렬을 나타낸다. 보이스피싱 탐지 정확도 및 재현율이 높게 유지된 반면, 일부 정상 음성에 대해 과탐지된 결과가 시각적으로 표현된다.

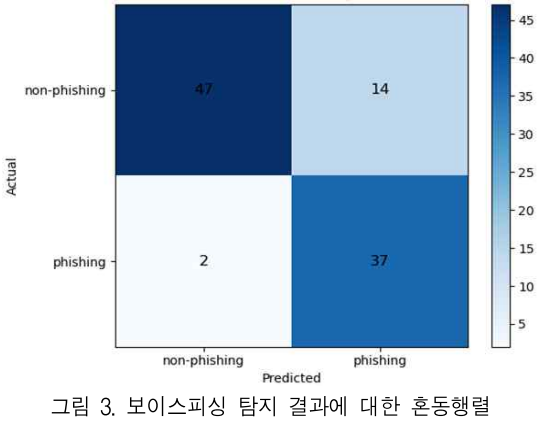

그림 3. 보이스피싱 탐지 결과에 대한 혼동행렬
Fig. 3. Confusion matrix of voice phishing detection results

개별 모델의 성능을 살펴보면, 텍스트 기반 KoBERT 분류기는 학습 데이터에 대해 정확도 99.8%, F1 점수 99.5%를 기록하였다. 음성 기반 CNN-LSTM 딥보이스 탐지 모델은 검증 데이터 기준 정확도 및 F1 점수 모두 100%로 매우 높은 성능을 달성하였다. 이는 CNN-LSTM 구조가 합성 음성과 실제 음성 간의 미세한 음향적 차이를 효과적으로 학습했음을 시사한다.

그림 4는 CNN-LSTM 기반 합성 음성 탐지 모델의 학습 과정에서 기록된 정확도와 F1 점수의 변화를 시각화한 결과이다. 전체적으로 학습 데이터와 검증 데이터 모두에서 정확도와 F1 점수가 높은 수준을 꾸준히 유지하였으며, 특히 2번째 에포크부터 10번째 에포크까지 모든 지표가 0.98 이상에서 안정적으로 수렴하였다. 학습 정확도와 F1 점수는 2번째 에포크에서 각각 0.98을 상회한 뒤 5번째 에포크 이후로는 1.00에 근접하며 거의 완전한 분류 성능을 달성하였음을 보여준다. 검증 데이터에서도 정확도와 F1 점수가 전 구간에 걸쳐 0.97~1.00 사이를 유지하였으며, 일시적으로 소폭 하락하는 구간(6번째 에포크 등)에서도 빠르게 회복되어 전반적으로 높은 일반화 성능을 나타내었다. 이러한 결과는 CNN-LSTM 모델이 합성 음성과 실제 음성의 특성을 효과적으로 학습했으며, 오버피팅 없이 안정적으로 분류 기능을 수행함을 시사한다. 특히, MFCC 기반 음성 특징 추출과 합성곱 신경망, 순환 신경망의 결합 구조가 합성 음성 탐지에 있어 매우 효과적임을 확인할 수 있다.

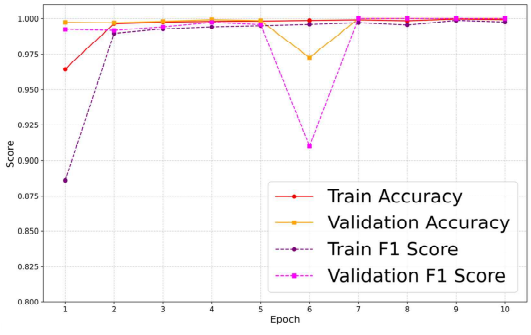

그림 4. CNN-LSTM 기반 음성 탐지 모델의 학습 과정
Fig. 4. Training process of the CNN-LSTM-based voice detection model

그림 5는 KoBERT 기반 텍스트 분류 모델의 학습 과정을 시각화한 결과이다. 학습 손실은 초기 에포크에서 급격히 감소하여 5번째 에포크 이후부터 1.0000로 수렴하였으며, 이는 모델이 빠르게 안정적인 표현 학습에 도달했음을 나타낸다. 학습 정확도와 F1 점수 역시 2번째 에포크에서 각각 0.9985, 0.9944로 도달한 이후 점차 상승하여, 5번째 에포크부터는 정확도 1.0000, F1 점수 1.0000으로 완전한 수렴을 달성하였다. 검증 데이터에 대한 성능도 안정적인 흐름을 보였다. 2번째 에포크에서 각각 0.9992, 0.9970의 높은 값을 기록한 뒤, 이후 에포크에서는 0.9996 이상의 정확도와 0.9985 수준의 F1 점수를 꾸준히 유지하였다. 이는 모델이 과적합 없이 일반화 성능도 안정적으로 확보했음을 의미하며, KoBERT의 Self-Attention 구조가 문장 내 단어 간 문맥 정보를 효과적으로 학습하여 텍스트 기반 보이스피싱 탐지에 적합함을 보여준다.

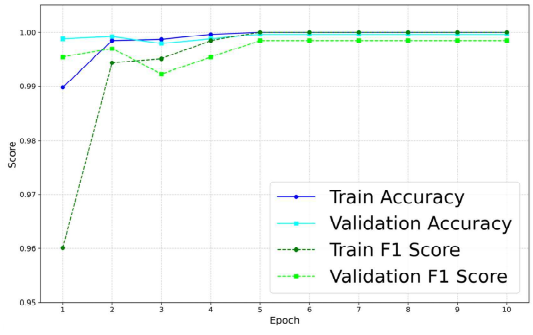

그림 5. KoBERT 기반 텍스트 분류 모델의 학습 과정
Fig. 5. Training process of the KoBERT-based text classification model

V. 결론 및 향후 과제

본 연구에서는 실제 통화 환경에서 발생하는 보이스피싱을 효과적으로 탐지하기 위해, 텍스트와 음성 정보를 병렬 분석하는 보이스 피싱 탐지 시스템을 제안하였다. 시스템은 KoBERT 기반 Self-Attention 텍스트 분류기와 CNN-LSTM 기반 딥보이스 판별기로 구성되며, Resemblyzer-KMeans 기반 화자 분리와 Whisper STT를 활용한 자동 전처리 구조를 갖춘다.

텍스트 모델은 금융감독원 실데이터 기반으로 높

은 분류 성능을 보였고, 음성 모델은 AI Hub 데이터를 기반으로 학습하여 합성 음성 판별에서 100% 정확도를 달성하였다. 최종 판단은 두 모델의 점수를 8:2 비율로 가중합하여 수행함으로써, 실제 환경에서도 높은 신뢰도로 탐지가 가능함을 입증하였다.

텍스트와 음성의 융합 비율(텍스트:음성 = 8:2)은 7:3, 8:2, 9:1의 조합을 비교한 결과, 전체 정확도와 피싱 탐지율을 동시에 최적화하는 값으로 선정되었다. 텍스트 기반 모델은 금융감독원 실데이터에서 95% 이상의 높은 정확도를 보였으며, 음성 기반 모델은 합성음성 판별에는 효과적이었지만 실사용 환경에서는 다소 낮은 신뢰도를 보여 텍스트 가중치를 높게 설정하였다.

실험 결과, 보이스피싱 시스템은 정확도 84.0%, 재현율 94.9%를 기록하며, 단일모달 기반 탐지의 한계를 효과적으로 극복함을 보여주었다. 특히 피싱 탐지 누락이 거의 없었다는 점은 시스템의 실용성과 신뢰도를 높인다. 다만, 짧은 대화나 맥락이 부족한 경우 일부 정상 대화가 과탐되는 사례가 있었으며, 향후 탐지 임계값 조정, 후처리 규칙, 가중치 조정 등을 통해 이러한 한계를 보완할 계획이다.

향후 연구 방향을 포함한 확장 가능성은 다음과 같다. 먼저, 화자 분리의 정밀도를 높이고 음성 구간 내 의미 변화까지 감지할 수 있도록, Whisper 기반의 다중 화자 diarization과 같은 고도화된 음성 처리 기술을 통합할 계획이다. 다음으로, 융합 비율을 자동으로 학습하여 상황에 따라 가중치를 동적으로 조정하는 방식을 적용하는 연구를 진행할 예정이다.

제안된 시스템은 최근 모바일 기기를 통한 음성·문자 기반 공격의 증가를 반영하여 설계되었다. 제안된 시스템은 모바일 앱이나 금융 솔루션에 연동되어, 통화 중 음성과 텍스트를 병렬 분석하고 실시간 경고 제공 또는 서버 전송을 통해 중앙 탐지 시스템과 연계될 수 있다. 또한 딥보이스·딥페이크 등 신종 공격에도 대응할 수 있도록, 음성 판별과 문맥 분석 기능을 통합적으로 제공한다.

모바일 환경에서는 '계좌', '송금', '수사' 등 핵심 단어에 선택적으로 높은 가중치를 부여함으로써 텍스트 분석 연산을 최적화하고, 토큰 사용량과 연산 비용을 절감하여 실시간 탐지가 가능하도록 설계되었다. 향후에는 모델 경량화, 온디바이스 추론, 클라우드 연동 등 다양한 구조 고도화를 통해 실용성과 확장성을 더욱 강화할 계획이다.

References

[1] D. Kim and J.-K. Sung, "Analysis on the status of phishing impersonating an administrative investigation agency and countermeasures", The Korean Journal of Public Safety and Information, Vol. 38, No. 2, pp. 91-120, Jun. 2024. https://doi.org/10.35147/knpsi.2024.38.2.91.

[2] K. Jung, Y. Kim, and Y. Min, "Application of Psychological Triggers to Voice Phishing : Focusing on Social Engineering", Journal of Social Science, Vol. 28, No. 4, pp. 181-194, Oct. 2017. https://doi.org/10.16881/jss.2017.10.28.4.181.

[3] D. Lee, "A Comparative Study of LLM- based Agents for Interactive Text-based Game Automation", The Journal of The Institute of Internet, Broadcasting and Communication, Vol. 25, No. 3, pp. 1-8, Apr. 2025. https://doi.org/10.7236/JIIBC.2025.25.3.1.

[4] J. Son and J. Song, "A Study on Cybercrime Detection Using Explainable AI Technique", The Journal of The Institute of Internet, Broadcasting and Communication, Vol. 25, No. 2, pp. 243-249, Apr. 2025. https://doi.org/10.7236/JIIBC.2025.25.2.243.

[5] M. Schmitt and I. Flechais, "Digital deception: Generative artificial intelligence in social engineering and phishing", Artificial Intelligence Review, Vol. 57, No. 324, Oct. 2024. https://doi.org/10.1007/s10462-024-10973-2.

[6] resemble-ai, "Resemblyzer: Voice embedding toolkit for speaker similarity and diarization", https://github.com/resemble-ai/Resemblyzer. [accessed: Jul. 23, 2025].

[7] A. Radford, J. W. Kim, and T. Xu, "Robust Speech Recognition via Large-Scale Weak Supervision", arXiv preprint, arXiv:2212.04356, Dec. 2022. https://doi.org/10.48550/arXiv.2212.04356.

[8] voicepaw, "so-vits-svc-fork: A realtime voice conversion system based on SoftVC VITS",

https://github.com/voicepaw/so-vits-svc-fork. [accessed: Jul. 23, 2025]

[9] National Information Society Agency (NIA), "Korean Conversational Speech Corpus", AI Hub, 2020. https://www.aihub.or.kr. [accessed: Jul. 23, 2025]

[10] M. K. M. Boussougou and D.-J. Park, "Attention-Based 1D CNN- BiLSTM Hybrid Model Enhanced with FastText Word Embedding for Korean Voice Phishing Detection", Mathematics, Vol. 11, No. 14. MDPIAG, pp. 3217, Jul. 2023. https://doi.org/10.3390/math11143217.

[11] B. Zhang, H. Cui, V. Nguyen, and M. Whitty, "Audio Deepfake Detection: What Has Been Achieved and What Lies Ahead", Sensors, Vol. 25, No. 7, pp. 1989, Apr. 2025. https://doi.org/10.3390/s25071989.

[12] A. Triantafyllopoulos, I. Tsangko, A. Gebhard, A. Mesaros, T. Virtanen, and B. Schuller, "Computer audition: From task-specific machine learning to foundation models", arXiv preprint, arXiv:2407.15672, Jul. 2024. https://doi.org/10.48550/arXiv.2407.15672.

[13] J. Kim, J. Kim, S. Wi, Y. Kim, and S. Son, "HearMeOut: Detecting Voice Phishing Activities in Android", Proc. of the 20th International Conference on Mobile Systems, Applications and Services (MobiSys '22), Portland, Oregon, pp. 422-435, Jun. 2022. https://doi.org/10.1145/3498361.3538939.

저자소개

김 지 원 (Jiwon Kim)

2020년 3월 ~ 현재 :
경상국립대학교 컴퓨터공학부
학사과정
관심분야 : 데이터 분석, 가상현실, 클라우드 컴퓨팅

구 슬 이 (Seuli Gu)

2022년 3월 ~ 현재 :
경상국립대학교 컴퓨터공학부
학사과정
관심분야 : 데이터 분석, 인공지능

김 혜 진 (Hyejin Kim)

2007년 8월 : 광주과학기술원
정보기전공학부(공학석사)
2014년 2월 : 광주과학기술원
정보통신학과(공학박사)
2014년 3월 ~ 2015년 6월 :
한국과학기술연구원
박사후연구원
2015년 7월 ~ 2020년 11월 : LG전자 CTO 선임연구원
2020년 12월 ~ 2023년 10월 : 서울기술연구원 수석연구원
2023년 11월 ~ 현재 : 서울연구원 연구위원
관심분야 : 증강현실, 인공지능, 빅데이터

강 창 구 (Changgu Kang)

2010년 2월 : 광주과학기술원
정보기전공학부(공학석사)
2017년 8월 : 광주과학기술원
전기전자컴퓨터공학부(공학박사)
2018년 3월 ~ 현재 :
경상국립대학교 컴퓨터공학부
부교수
관심분야 : 컴퓨터 그래픽스, 증강현실, 인공지능