JDFR Journal of Data Forensics Research
데이터포렌식연구, 1권 1호, 2024년 12월, 177-183면
Journal of Data Forensics Research, Volume 1, Issue 1, December 2024, Pages 177-183.
https://doi.org/10.12972/JDFR.2024.1.1.11
Print ISSN 3058-9460
Online ISSN 3058-9479

기고문

LLM과 온디바이스 AI 기반의 보이스피싱 탐지 기술의 미래

박현호

한국전자통신연구원 선임연구원

교신저자: 박현호, hyunhopark@etri.re.kr

요약

보이스피싱은 피해자의 개인정보를 탈취하거나 금전적 피해를 초래하는 음성 기반 사기 행위로, 수법이 점차 정교화되고 다양화되고 있다. 특히 딥보이스, 딥페이크, LLM(Large Language Model)과 같은 최신 기술이 악용되면서 피해자가 이를 감지하기 더욱 어려워지고 있다. 이에 따라 보이스피싱 탐지 연구는 기존의 통계 기반 방식에서 딥러닝 기반 자연어 처리 기법으로 발전하고 있으며, LLM을 활용한 탐지가 주목받고 있다. LLM은 대규모 데이터를 학습하여 보이스피싱 메시지와 정상 메시지의 미세한 차이를 정교하게 탐지할 수 있는 능력을 지니고 있으며, 온디바이스 AI는 네트워크 연결 없이도 실시간으로 메시지를 감지하며 개인정보를 보호할 수 있는 장점이 있다. 결론적으로, LLM과 온디바이스 AI의 결합은 다양한 신종 보이스피싱 수법에 대응할 수 있는 강력한 도구로 자리 잡고 있다. 이러한 기술은 보이스피싱을 사전에 탐지하고 차단하며, 개인정보 보호와 실시간 대응이 가능한 효과적인 방어 체계를 제공한다. 지속적으로 발전하는 기술을 활용해 더 효과적인 대응 방안을 마련하는 것이 중요하다.

주제어

보이스피싱, 딥보이스, 딥페이크, LLM, 온디바이스 AI

Received: November 20, 2024
Accepted: December 31, 2024
Published: December 31, 2024

Invited Article

Future of Voice Phishing Detection Technology Based on LLM and On-Device AI

Hyunho Park

Senior Researcher, Electronics and Telecommunications of Research Institute, Republic of Korea

Corresponding Author: Hyunho Park, hyunhopark@etri.re.kr

ABSTRACT

Voice phishing is a voice-based fraud activity that aims to steal victims' personal information or cause financial losses. Its methods are becoming increasingly sophisticated and diverse. In particular, advanced technologies such as Deep Voice, Deep Fake, and LLM (Large Language Models) are being exploited, making it even more challenging for victims to detect these attacks. Consequently, research on voice phishing detection has evolved from traditional statistical methods to deep learning-based natural language processing techniques, with a focus on utilizing LLMs. LLMs, trained on large-scale datasets, possess the ability to meticulously detect subtle differences between voice phishing messages and legitimate messages. On-device AI, on the other hand, offers the advantage of real-time message detection without requiring network connectivity, ensuring enhanced privacy protection. In conclusion, the combination of LLMs and on-device AI has emerged as a powerful tool to counter various new types of voice phishing attacks. These technologies enable proactive detection and prevention of voice phishing, providing an effective defense system that ensures real-time response and privacy protection. It is crucial to leverage continuously advancing technologies to develop more effective countermeasures in the future.

KEYWORDS

Voice Phishing, Deep Voice, Deepfake, LLM, On-Device AI

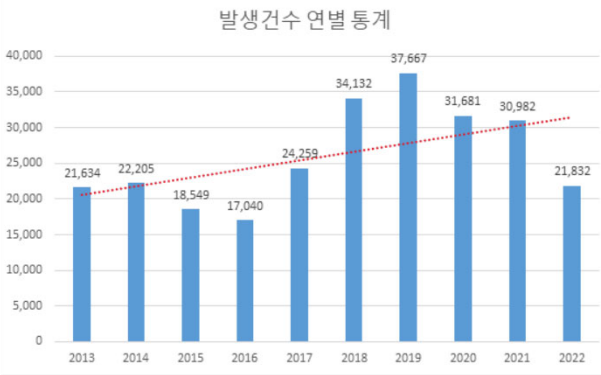
1. 서론

보이스피싱(Voice Phishing)은 가해자가 피해자와 음성 통화를 통해 얻은 개인정보를 이용해 금융 정보를 탈취하거나 금전적 피해를 입히는 사기 행위이다[1]. <Figure 1>에 나타난 바와 같이 보이스피싱 발생 건수는 연도별로 다소 변동이 있었으나, 장기적으로는 꾸준히 증가하고 있으며, <Figure 2>에 따르면 보이스피싱으로 인한 피해 금액도 지속적으로 증가해 이에 따른 경제적 손실과 심리적 피해가 심각한 수준에 이르고 있다[2].

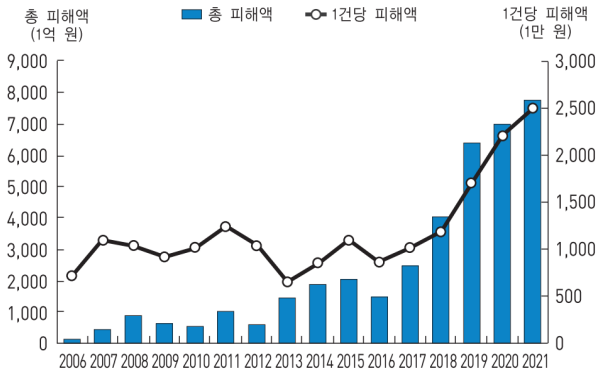
보이스피싱의 유형은 주로 은행 등 금융회사를 사칭해 대출 관련 사기를 시도하는 "대출사기형"과 검찰이나 경찰 등 정부기관을 사칭해 피해자를 속이는 "기관사칭형"으로 나뉜다. 그러나 최근 보이스피싱 유형과 수법이 다양화되면서, 피해자가 이를 감지하고 대응하기가 점점 어려워지고 있다.

보이스피싱 감지 및 대응을 위한 연구는 초기의 통계 기반 패턴 인식이나 규칙 기반 탐지에서 딥러닝 모델을 활용한 자연어 처리 기법으로 발전하고 있다. 특히, BERT(Bidirectional Encoder Representations from Transformers)와 같은 사전 학습(pre-trained) 언어 모델을 활용해 음성 데이터를 텍스트로 변환한 후, 텍스트 내 특정 키워드와 문맥을 분석하여 보이스피싱 여부를 판단하는 방식이 주목받고 있다. 예를 들어, 연구[3]에서는 사전 학습된 한국어 모델 KoBERT를 기반으로 한국어 보이스피싱 데이터셋을 활용해 보이스피싱 데이터를 분류하였다. 또한, 연구[4]는 통화 음성을 녹음 및 텍스트로 변환한 후 KoBERT를 이용해 보이스피싱 패턴을 학습시키고, 실시간으로 보이스피싱을 감지하는 시스템을 설계하였다.

최근 1,000억 개 이상의 학습 파라미터를 가진 대형 언어 모델(LLM, Large Language Model)이 보이스피싱 탐지 연구에서 중요한 기술로 주목받고 있다. LLM은 대규모 데이터를 학습하여 보이스피싱 메시지의 복잡한 언어 패턴을 정교하게 분석할 수 있는 역량을 갖추고 있으며, 온디바이스 AI는 텍스트와 음성을 실시간으로 처리하면서 사용자 개인정보를 보호할 수 있는 잠재력을 지니고 있다. 본 기고문에서는 LLM과 온디바이스 AI를 활용한 보이스피싱 탐지 기술의 최신 연구를 검토하고, 이러한 기술이 기존 감지 방식보다 어떤 점에서 뛰어난 성능을 제공하는지 논의한다. 또한, 음성 및 영상 합성 기술의 발전으로 예상되는 신종 보이스피싱 수법과 이에 대응하기 위한 연구 방향도 함께 제시하고자 한다.



<Figure 1> 연도별 보이스피싱 발생건수 추이 [1]



<Figure 2> 연도별 보이스피싱 피해액 [2]

2. 보이스피싱 범죄 유형

금융감독원은 보이스피싱의 주요 유형을 다음과 같이 10가지로 분류하였다[5]. 이 분류에는 음성 통화뿐만 아니라 메신저나 문자 메시지를 이용한 개인정보 유출 사례도 포함된다. 이는 스마트폰 보급과 통신 환경의 데이터 통신 중심 변화에 따라 보이스피싱, 스미싱(smishing), 악성 앱 설치가 혼재된 범죄가 발생하고 있는 현상을 반영한다.

1. 자녀 납치 및 사고 빙자 편취: 자녀나 가족의 사고나 납치를 가장하여 긴급 송금을 요구하는 방식으로, 피해자의 가족에 대한 애착과 심리적 취약점을 이용한다.
2. 메신저상에서 지인을 사칭하여 송금요구: 메신저 계정을 해킹한 후 가족이나 지인을 사칭하여 긴급 상황을 연출하고 송금을 유도한다.
3. 인터넷 뱅킹을 이용해 카드론 대금 및 예금 등을 편취: 피해자의 인터넷 뱅킹 정보를 탈취하거나 가짜 사이트를 통해 계좌에서 자금을 이체하는 방식으로, 카드번호와 비밀번호 등 민감한 정보를 확보하여 금융 피해를 발생시킨다.
4. 금융기관 명의의 허위 긴급공지 문자로 기망, 피싱사이트로 유도, 예금 등 편취: 은행이나 금융기관을 사칭하여 긴급 문자를 보내고, 피해자가 이를 신뢰하여 금융 정보를 제공하거나 대출을 유도받아 자금을 송금하게 한다.
5. 전화통화를 통해 텔레뱅킹 이용정보를 알아내어 금전 편취: 주로 고령층을 대상으로 텔레뱅킹 가입 확인이나 개인정보 확인을 빙자하여 텔레뱅킹 정보를 확보한 후 계좌를 악용하여 자금을 유출한다.
6. 피해자를 기망하여 자동화기기로 유인 편취: 피해자를 ATM 등 자동화 기기로 유도하여 세금 환급이나 공공기관 수수료 등의 명목으로 자금을 송금하게 한다.
7. 피해자를 기망하여 피해자에게 자금을 이체하도록 하여 편취: 경찰, 금융감독원, 공공기관 등을 사칭하여 거래내역 확인이나 계좌 안전 확인 등을 빙자하여 피해자가 예금을 인출하고 사기계좌로 이체하도록 유도한다.
8. 신용카드 정보를 취득후 ARS를 이용한 카드론 대금 편취: 명의도용과 정보유출을 통해 신용카드 정보를 확보한 후 ARS나 인터넷을 통해 카드론 대금 상환 절차를 안내하며 송금을 요구하여 자금을 편취한다.
9. 상황극 연출에 의한 피해자 기망 편취: 경찰, 검찰, 수사기관 등을 사칭하여 피해자가 범죄에 연루되었다고 주장하며 금전 송금을 요구한다.
10. 물품대금 오류송금 빙자로 피해자를 기망하여 편취: 문자 메시지나 전화를 통해 물품 대

금이나 택배 관련 착오를 가장하여 피해자에게 연락을 유도한 뒤 송금을 요구하거나 잘못된 계좌번호로 입금을 요청한다.

이처럼 보이스피싱은 피해자의 신뢰를 얻거나 심리적 취약점을 자극하는 다양한 수법으로 이루어진다. 이러한 세분화된 유형별 분석은 효과적인 감지 모델 개발과 대응 방안 연구에 중요한 기초 자료가 된다.

3. 신종 보이스피싱 수법

보이스피싱 수법은 인공지능 기술의 발전과 함께 더욱 정교해지고 있다. 특히 딥보이스(DeepVoice), 딥페이크(DeepFake), LLM과 같은 기술이 악용되며 새로운 형태의 보이스피싱 사례가 발생하고 있다

1. 딥보이스를 활용한 보이스피싱: 딥보이스는 딥러닝 기술을 통해 특정인의 목소리를 정교하게 모방하는 기술로, 이를 활용하여 가족이나 지인의 목소리를 흉내 내어 피해자를 속이는 수법이 등장하고 있다[6]. 예를 들어, 2021년 UAE의 한 은행은 딥보이스 기술로 임원의 목소리를 흉내 낸 전화사기단에게 3,500만 달러를 송금하며 큰 피해를 입었다. 국내에서도 2023년, 음성 변조 애플리케이션을 이용해 일본인 가수, 그의 동료, 소속사 팀장의 목소리를 모방하여 피해자를 속이고 1,600여만 원의 현금을 갈취한 사례가 보고되었다.
2. 딥페이크를 활용한 사기: 딥페이크 기술은 특정인의 얼굴과 목소리를 동시에 모방할 수 있어 영상 통화나 동영상 메시지를 통해 피해자를 속이는 수법으로 발전하고 있다. 2023년 6월, 국내에서 딥보이스와 딥페이크 기술을 결합해 유명 검사를 사칭하며 피해자의 신상 정보를 갈취하려는 시도가 있었다[6]. 이러한 기술의 발전은 보이스피싱의 설득력을 크게 강화하고 있다.
3. LLM을 활용한 피싱 메시지 생성: LLM은 대규모 데이터를 학습하여 실시간으로 대화를 조작하는 수법이 예상된다. 연구[7]에 따르면, LLM 기반 모델(GPT-3, GPT-3.5, GPT-4)을 이용해 특정 개인, 조직, 기업을 대상으로 하는 맞춤형 피싱 메시지를 생성한 결과, 모델의 버전이 발전할수록 더욱 자연스럽고 유창한 피싱 메시지가 만들어지는 것으로 확인되었다.

보이스피싱 범죄자가 LLM을 딥보이스 및 딥페이크와 함께 사용한다면, 피해자가 이를 인지하기 더욱 어려워질 가능성이 높다. 이러한 신종 보이스피싱 수법은 기존의 탐지 방식을 무력화할 수 있어 효과적인 대응 기술 개발이 시급하다.

4. 보이스피싱 탐지 기술의 미래: LLM과 온디바이스 AI의 활용

보이스피싱 탐지를 위해 LLM과 온디바이스(On-Device) AI의 활용이 필수적이다. 이는 기술 발전과 더불어 보이스피싱 범죄가 점차 정교화됨에 따라 기존의 단순 패턴 인식 방식으로는 다양한 공격 유형과 고도화된 사기 수법에 대응하기 어려운 한계를 극복하기 위해 필요하다. 특히 개인화된 보이스피싱 메시지와 음성 합성 기술이 악용되는 사례가 늘어나면서, 실시간 보이스피싱 탐지를 가능하게 할 LLM과 온디바이스 AI의 역할이 강조되고 있다.

먼저, LLM은 대규모 언어 데이터를 학습하여 피싱 메시지의 복잡한 언어적 패턴을 분석하는 데 탁월한 성능을 보인다[8]. LLM은 정상적인 대화와 사기성 대화의 미세한 차이를 감지하고, 문맥 내 의심스러운 단어와 문구를 파악함으로써 기존의 규칙 기반 탐지보다 우수한 성능을 제

공한다. 또한, 지속적으로 진화하는 보이스피싱 수법에 유연하게 대응하기 위해 새로운 언어 패턴과 트렌드를 학습함으로써 탐지 모델의 신뢰성을 유지할 수 있다.

온디바이스 AI는 데이터 처리를 클라우드로 전송하지 않고 사용자 기기에서 직접 실행할 수 있어, 실시간 감지와 개인정보 보호 측면에서 매우 유용하다[9]. 이는 사용자의 민감한 데이터가 외부 서버로 전송되지 않음으로써 개인정보 보호가 강화되고, 네트워크 연결 상태에 관계없이 독립적으로 작동할 수 있어 오프라인 환경에서도 보이스피싱 탐지가 가능하다. 음성 메시지를 실시간으로 분석하고 위험이 감지될 경우 즉시 경고를 제공함으로써 피해를 예방할 수 있다.

LLM과 온디바이스 AI를 결합한 시스템은 텍스트와 음성을 함께 분석하는 멀티모달 감지를 가능하게 하며, 특히 가족이나 주변 지인의 목소리나 말투를 사칭하는 보이스피싱에 효과적이다. 온디바이스 AI는 텍스트와 음성 간의 불일치뿐만 아니라 말투, 억양, 감정 변화와 같은 음성 특성을 분석하여 사기 여부를 판단하며, LLM은 대화 스타일이나 말투를 학습해 사칭형 보이스피싱 메시지를 정확히 식별한다. 이를 통해 실시간으로 맞춤형 경고를 제공하며 사칭 시도를 효과적으로 차단할 수 있다.

LLM과 온디바이스 AI는 사용자 피드백을 바탕으로 새로운 보이스피싱 패턴을 지속적으로 학습하여 탐지 능력을 강화할 수 있다. 사용자가 의심스러운 메시지나 통화를 표시하면, 온디바이스 AI는 이를 학습하고 이후 유사한 사례를 신속히 탐지할 수 있는 환경을 구축할 수 있다. 이는 보이스피싱 수법의 진화에 대응해 모델의 신뢰성과 탐지 성능을 유지하는 데 중요한 역할을 한다.

결론적으로, LLM과 온디바이스 AI는 보이스피싱 탐지 체계를 강화하는 데 핵심적인 역할을 한다. LLM의 고도화된 언어 분석 능력과 온디바이스 AI의 실시간 처리 및 개인정보 보호 기능은 서로 보완적으로 작용하여 효과적인 방어 체계를 구축할 수 있다. 보이스피싱 수법이 더욱 정교해질 미래를 대비하기 위해 이들 기술의 결합은 신뢰할 수 있는 대응 시스템을 제공하는 데 필수적이다.

5. 결론

보이스피싱은 기술의 발전과 함께 수법이 점점 다양화되고 정교해지면서 사회적으로 큰 위협으로 자리 잡고 있다. 먼저, 기존의 보이스피싱 범죄 유형을 살펴보면, 피해자의 심리적 취약점을 이용한 다양한 접근 방식이 활용되고 있음을 확인할 수 있다. 이러한 유형은 보이스피싱이 단순한 사기 행위를 넘어 점점 더 체계적이고 조직화된 범죄로 진화하고 있음을 보여준다. 또한, 딥보이스, 딥페이크, LLM과 같은 최신 기술이 보이스피싱에 악용되는 사례를 조명하였다. 딥러닝 기술을 활용한 목소리 모방, 영상 조작, 그리고 LLM 기반의 자연스러운 피싱 메시지 생성은 피해자가 보이스피싱을 인지하기 어렵게 만들며, 기존의 탐지 및 대응 방식을 무력화할 가능성을 보여준다. 끝으로, LLM과 온디바이스 AI의 활용 방안에서는 이러한 신종 수법에 대응하기 위한 기술적 접근이 논의되었다. LLM을 활용한 정교한 언어 분석 및 탐지와 온디바이스 AI를 통한 실시간 감지 및 개인정보 보호는 미래의 보이스피싱 대응에서 핵심적인 역할을 할 것이다. 특히, 피해자의 데이터를 보호하면서 오프라인 환경에서도 실시간으로 대응할 수 있는 기술 개발은 반드시 필요한 과제이다.

결론적으로, 보이스피싱에 효과적으로 대응하기 위해서는 지속적으로 발전하는 기술에 대한 다각적인 접근이 요구된다. 범죄 유형과 신종 수법에 대한 깊은 이해를 기반으로 LLM과 온디바이스 AI와 같은 최신 기술을 효과적으로 활용한다면, 이러한 범죄를 사전에 탐지하고 차단하는 데 중요한 기여를 할 수 있을 것이다.

사사(Acknowledgements)

본 기고문은 한국전자통신연구원 내부연구개발사업 “보이스피싱 탐지를 위한 디바이스 자율 탐색 및 최적화 기반 초경량 온디바이스AI 알고리즘 연구”(24BR1200, 24RR1300) 지원으로 작성되었습니다.

참고문헌 (References)

[1] Seo JB. 2022. Current Status, Types, Trends, and Implications for Countermeasures Against Voice Phishing. Korean Social Trends, 307-315.

[2] Jung JY, Yeom YH. 2023. Analyzing the Trend of Voice Phishing to Identify Issues and Derive Countermeasures. Criminal Investigation Institute, 9(2), 153-157.

[3] Boussougou MKM, Park DJ. 2022. Exploiting Korean Language Model to Improve Korean Voice Phishing Detection. The Transactions of the Korea Information Processing Society/Software and Data Engineering, 11(10), 437-446.

[4] Kim YJ, Lee BY, Kang AR. 2024. Design of Real-Time Voice Phishing Detection Techniques using KoBERT, Korea Society of Computer and Information, 32(1), 95-96.

[5] FSS (Financial Supervisory Service). 2024. Voice Phishing Guardian - Major Fraud Types. Seoul, Korea: FSS. Available at: https://fss.or.kr/fss/main/contents.do?menuNo=200565

[6] Kim SH, Park KS. 2024. [Exclusive] TV prosecutor’s face and voice were deepfake voice phishing ([단독] TV 나온 그 검사의 얼굴·목소리, 보이스피싱 딥페이크였다). Available at: https://www.seoul.co.kr/news/society/law/2024/04/02/20240402001006 accessed on 2024.04.02.

[7] Hazell J. 2023. Large Language Models Can Be Used To Effectively Scale Spear Phishing Campaigns. arXiv preprint arXiv:2305.06972.

[8] Chataut R, Gyawali, Usman Y. 2024. Can ai keep you safe? a study of large language models for phishing detection. Proceedings of 2024 IEEE 14th Annual Computing and Communication Workshop and Conference, Las Vegas, 0548-0554.

[9] Seo JW, Lee JS, Kim HW, et al. 2024. On-Device Smishing Classifier Resistant to Text Evasion Attack. IEEE Access, 12, 4762-4779.