

보이스피싱 탐지 AI 강화를 위한 ATT&CK - RAG 프레임워크

설동명, 김지용
한국전자통신연구원

dmsul@etri.re.kr, kji@etri.re.kr

ATT&CK-RAG Framework for Strengthening AI for Voice Phishing Detection

Seol Dong Myung, Kim Ji Yong
Electronics and Telecommunications Research Institute.

요약

본 논문은 보이스피싱의 사회공학적 특성을 체계적으로 분석하기 위해 설계된 보이스피싱 ATT&CK(Adversarial Tactics, Techniques, and Common Knowledge) 프레임워크를 제안한다. 본 프레임워크는 최근 급증한 AI 음성 모방, 정부지원 빙자 대출, 가상자산 투자 사기 등 최신 사례를 기반으로, 공격 전술(Tactics)과 기술(Techniques)을 명확히 분류한다. 특히 피해자 중심의 관점에서 심리적 조작 과정을 구조화하여, 효과적인 예방·대응 전략 수립에 기여한다. 또한, 프레임워크를 활용한 RAG(Retrieval-Augmented Generation) 기반 시나리오 자동 생성 방법론을 제시함으로써, 변종 및 신종 보이스피싱 탐지 AI 학습 데이터 확장과 성능 향상을 동시에 달성한다. 제안한 접근법은 기관과 개인의 능동적 위협 대응을 지원하며, 향후 보안 교육·실시간 탐지 시스템에 활용 가능하다.

I. 서론

보이스피싱은 단순한 금전적 피해를 넘어 사회 전반에 심각한 신뢰 훼손을 초래하는 범죄로 발전하고 있다[1]. 최근 공격자는 발신 번호 조작, AI 음성 합성, 맞춤형 사회공학 기법 등 첨단 기술을 적극 활용하고 있다. 그러나 기존 사이버 보안 위협 모델링은 주로 네트워크·시스템 취약점에 초점을 맞춰, 인간 심리의 취약성을 악용하는 보이스피싱 특성을 충분히 반영하지 못했다.

이에 본 연구는 피해자 관점의 심리·행동 기반 위협 모델인 보이스피싱 ATT&CK 프레임워크를 설계하였다. 또한 해당 프레임워크를 RAG 기법과 결합해 변종 시나리오를 자동 생성함으로써, 실제 탐지 AI의 학습 데이터 부족 문제를 해결하고자 한다.

II. 보이스피싱 ATT&CK 프레임워크

보이스피싱 범죄는 발신자 번호 조작, 정교한 시나리오 개발, AI 음성 합성 등 기술적 진보와 결합하여 끊임없이 고도화되고 있다. 특히 코로나 19, 고금리 등 사회적 변화를 악용한 수법들이 급증하며, 피해가 더욱 광범위해지고 있다. 이러한 공격 방식은 IT 시스템의 취약점보다 인간의 심리적 취약점에 의존하는 사회공학적 기법을 핵심으로 한다. 따라서 단순히 기술적 방어에 집중하는 기존 모델로는 보이스피싱 공격을 효과적으로 방어하기 어렵다. 본 프레임워크는 이와 같은

배경에서 보이스피싱 공격의 전술과 기술을 표준화된 방식으로 분류하고 공유함으로써, 대중 및 기관의 위협 이해도를 높이고 피해를 최소화하는 것을 목표로 한다.

2.1 보이스 피싱 ATT&CK 프레임워크의 구성 요소

보이스피싱 ATT&CK 프레임워크는 기존 IT 중심의 ATT&CK과 차별화되는 고유한 구성 요소를 포함한다. 이 구성 요소들은 보이스피싱의 비기술적이고 심리적인 복잡성을 포착한다.

2.1.1 전술 (Tactics)

프레임워크는 공격의 목적에 따라 6 가지 주요 전술을 정의한다. '초기 접촉 및 접근'을 시작으로, 피해자의 경계심을 허무는 '신뢰 구축', 비합리적 판단을 유도하는 '심리적 압박 및 조작', 민감 정보를 탈취하는 '정보 획득', 금전적 이득을 취하는 '금전 편취', 그리고 추적을 어렵게 만드는 '흔적 제거 및 회피'로 구성된다. 이 전술들은 기술적 시스템이 아닌 '인간'이라는 가장 취약한 고리를 목표로 한다는 보이스피싱의 본질을 반영한다.

2.1.2 기술 (Techniques)

각 전술 아래에는 특정 행동을 수행하기 위한 구체적인 방법인 기술이 정의되어 있다. 프레임워크는 2020 년 이후의 최신 사례를 반영하여 다음 기술들을 포함한다.

- 초기 접촉: 전화 발신(T1001.001), 문자/메시징 앱(T1001.002)을 통한 접근과 함께, AI 음성

모방(T1004.005)과 같은 첨단 기술을 활용한 지인 사칭(T1002.002) 등이 포함된다.

- 정보 획득: 피해자의 개인 식별 정보(T1004.001) 및 금융 정보(T1004.002)를 직접 요구하는 것 외에도, 악성 앱 설치 유도(T1004.003)나 가짜 웹사이트 유도(T1004.004)를 통해 정보를 탈취하는 기술이 상세히 명시된다.

- 금전 편취: 단순히 계좌 이체(T1005.001)를 유도하는 것을 넘어, 피해자 명의로 고액 대출을 실행하고 편취(T1005.004)하거나 비대면 계좌를 개설하여 악용(T1005.005)하는 최신 수법들이 포함되어 있다.

2.2 보이스 피싱 ATT&CK 프레임워크 설계

본 프레임워크는 다음의 핵심 철학을 기반으로 설계되었다. 첫째, 피해자 중심 관점이다. 공격자의 기술적 침투 경로보다는 피해자가 겪는 인지적, 감정적 과정에 초점을 맞추므로써, 실제 피해 상황을 더 잘 이해할 수 있도록 돕는다. 둘째, 실제 사례 기반이다. 금융감독원 등 공신력 있는 기관의 발표 자료와 2020 년 이후의 최신 사례를 면밀히 분석하여 프레임워크의 실용성과 신뢰성을 확보했다. 셋째, 예방 및 교육 중점이다. 공격 전술과 기술을 명확히 이해함으로써 효과적인 예방 교육 프로그램을 개발하고, 개인의 보안 의식을 높이는 데 활용될 수 있도록 했다. 마지막으로 확장성 및 유연성을 고려하여 새로운 보이스피싱 수법에 유연하게 대응할 수 있도록 했다.[2]

III. 보이스피싱 탐지 AI 를 위한 RAG 기반 시나리오 생성

보이스피싱 탐지 AI 의 성능은 방대한 양의 고품질 학습 데이터에 크게 의존한다. 그러나 기존의 실제 사례 데이터만으로는 끊임없이 진화하는 변종 및 신종 공격 시나리오에 효과적으로 대응하기 어렵다. 이에 본 논문은 보이스피싱 ATT&CK 프레임워크를 활용하여 RAG(Retrieval-Augmented Generation) 기반으로 다양한 시나리오를 자동으로 생성하는 방법론을 제안한다.[3]

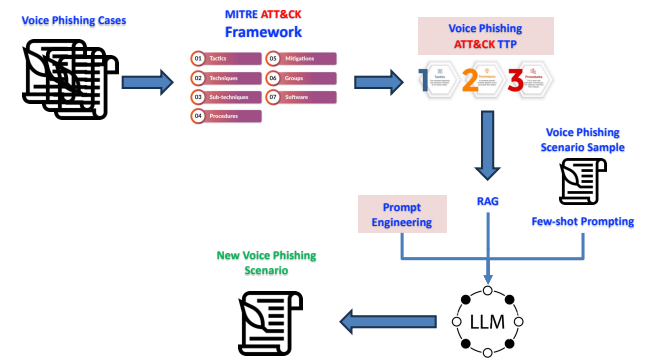
3.1 RAG 기반 시나리오 생성 방법

RAG 모델은 '검색(Retrieval)'과 '생성(Generation)'의 두 가지 핵심 단계로 구성된다.

- 검색(Retrieval): 보이스피싱 ATT&CK 프레임워크의 전술 및 기술 데이터베이스, 실제 보이스피싱 대본, 그리고 보이스피싱에 사용될 수 있는 일상 대화 패턴 등 다양한 정보 소스를 벡터화하여 지식 기반을 구축한다. 특정 시나리오를 생성하고자 할 때, 사용자는 결합하고자 하는 전술(예: T1004.005 AI 음성 모방)과 기술(예: T1005.004 대출 실행)을 프롬프트로 입력한다. 시스템은 이 프롬프트와 가장 연관성이 높은 정보(예: T1004.005 와 T1005.004 에 대한 설명, 관련 실제 대화 예시)를 지식 기반에서 검색하여 가져온다.

- 생성(Generation): 검색된 정보는 생성 모델(Generative Model), 즉 LLM(Large Language Model)의 입력으로 들어간다. LLM 은 이 정보를 바탕으로 특정 전술과 기술이 결합된, 현실적이고 자연스러운 대화 스크립트를 생성한다. 이 과정에서

다양한 화법, 말투, 상황(예: 자녀 빙자, 금융 기관 사칭 등)을 적용하여 다채로운 변종 시나리오를 만들어낼 수 있다.



[신종 보이스 피싱 시나리오 생성]

3.2 RAG 기반 시나리오의 장점

이 방법론은 다음과 같은 이점을 가진다. 첫째, 데이터 희소성 문제 해결: 실제 데이터의 한계를 넘어, 무한에 가까운 변종 시나리오를 생성하여 AI 학습 데이터의 양을 극적으로 늘릴 수 있다. 둘째, 실용성과 다양성 확보: 프레임워크의 전술과 기술을 기반으로 생성되므로 현실성이 높고, 다양한 조합을 통해 공격의 복잡성을 반영한 시나리오를 만들어낼 수 있다. 셋째, 능동적 대응: 새로운 공격 트렌드가 발견되면 해당 정보를 지식 기반에 추가하여, 즉시 이를 반영한 시나리오를 생성하고 탐지 모델을 업데이트할 수 있다.

IV. 결론

본 논문에서는 보이스피싱의 사회 공학적 특성을 체계적으로 분석한 보이스피싱 ATT&CK 프레임워크를 제안하고, 나아가 이 프레임워크를 활용하여 보이스피싱 탐지 AI 의 성능을 혁신적으로 향상시킬 수 있는 RAG 기반 시나리오 생성 방법론을 제시했다. 이 프레임워크는 보이스피싱 예방 교육 뿐만 아니라, AI 기반 보안 시스템 개발의 핵심 도구로 활용될 수 있다. RAG 기반 생성 모델은 기존 데이터의 한계를 극복하고, 현실적이고 다양한 변종 시나리오를 만들어냄으로써 AI 모델이 미래의 예측 불가능한 위협에 더욱 효과적으로 대응할 수 있도록 돕는다.

ACKNOWLEDGMENT

이 논문은 2025 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.RS-2025-02215393, (2 세부) 알려지지 않은 신종 보이스피싱 탐지·예측 기술개발).

참 고 문 헌

[1] 금융감독원, “보이스피싱 피해 현황 및 예방 대책,” 2024.

[2] MITRE ATT&CK, <https://attack.mitre.org/>

[3] AWS, “Retrieval-Augmented Generation,” <https://aws.amazon.com/ko/what-is/retrieval-augmented-generation/>