AI 기반 피싱 공격의 현황과 대응 방안

딥페이크와 생성형 AI의 위협

목차

AI기반 피싱이란

- AI(특히 생성형 AI•딥페이크 기술)을 이용해 타인을 속이고 민감정보•금전을 탈취하는 공격
- 기존 피싱과의 대비 차이점

기존 피싱	AI 기반 피싱
오타,부자연스러운 표현	완벽한 문법,맞춤형 메세지
일반적인 메일,문자	딥페이크 영상,음성,챗봇
대량 동일 메시지	맞춤형(스피어 피싱)자동 생성

주요 특징

1. 딥페이크 음성•영상 활용
   - 인물의 목소리, 얼굴을 AI로 복제하여 실시간 또는 사전 제작 영상으로 사칭
   - 1~2분 분량의 음성 샘플만 있으면 고품질 복제가 가능하다.
   - 영상 합성 또는 리얼타임 페이스 스왑 기술로 구현
2. 자연스러운 문장 생성
   - 직책, 관심사에 맞춘 맞춤형 문장 생성
   - 설득형 메시지 또는 심리적 압박 문구 등을 자동 생성
   - 생성형 AI로 오타, 문법 오류 없는 피싱 문자 작성
3. 타깃 맞춤형 공격(스피어 피싱)
   - 오픈소스로 수집한 정보(SNS, 뉴스, 사내 조직도 등))분석
   - AI가 대상을 프로파일링 → 개인별 맞춤형 공격 콘텐츠 생성
   - 자동화 스크립트를 이용해 수백 명에게 각각 다른 내용의 피싱 메시지 발송

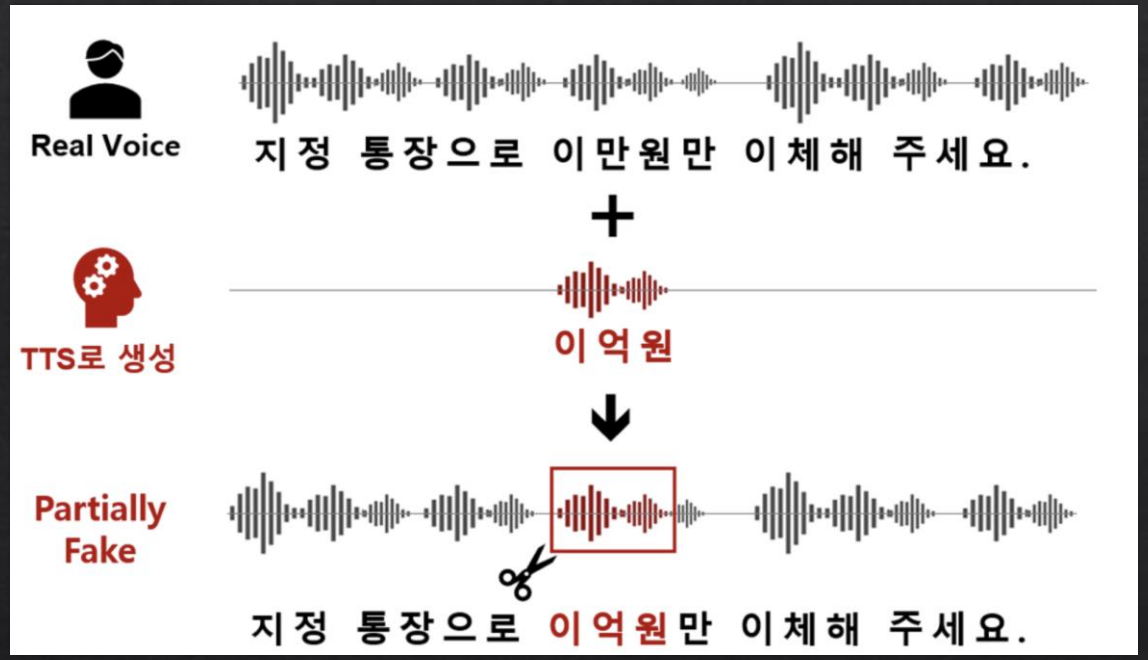
홍콩 딥페이크 화상회의 사건

- 내용: 회사의 최고재무책임자(CFO)에게 거액의 돈을 송금하라는 메일을 받음 피싱이라고 생각하여 의심
  하지만 화상 회의에서 같은 지시를 받자 의심을 거두고 송금
  AI로 CFO등 여러 임원 얼굴,음성 복제 → 회계 담당자가 총 HK$200M (약 2,500만 USD) 송금
  이체가 이뤄지기까지 피해 직원의 메신저와 이메일, 화상 통화로 계속 연락했다.

대응 방안

- 콜백 확인 프로토콜 설정
- 딥페이크 탐지 도구 도입
  행안부,국과수 개발중 데이터 231만 건 딥러닝 시켜 가짜 여부, 확률 도출
  영상은 94.98% 음성은 86.20%

부분 변조 공격



부분 변조 공격
음성의 생성 과정을 보여 주고 있는데
타깃의 실제 음성에서 중요 부분에 TTs로
합성한 음성 일부분을 대체하여 생성한다.
부분 음성은 대부분의 음성이 사람
음성이고 일부분만 디페이크 음성이므로
사람이 인지하기 어렵고 탐지 시스템도
탐지하기 어렵다.
그림과 같이 주요 정보를 변경하면 발화의
내용이 전혀 다른 내용으로 변경되므로
공격의 효과를 그대화 할수 있다.