Whisper와 LLaMA 기반 딥보이스 보이스 피싱 방지 앱

이건호*, 이슬희*, 이태규*, 이호진*, 김은도°, 전승현*

대전대학교 컴퓨터공학과*, KT°

eummm1215@gmail.com*, iseulhui666@gmail.com*, ltk5269@naver.com*, Leeho2273@naver.com*, eundo.kim@kt.com°, creemur@dju.kr*

A Deep Voice Phishing Prevention Application Based on Whisper and LLaMA

Gun Ho Lee*, Seul Hee Lee*, Tea Kyu Lee*, Ho Jin Lee*, Eun-Do Kim°, Seung Hyun Jeon*

요 약

본 논문은 딥보이스 기술을 활용한 보이스 피싱을 탐지하기 위한 AI 기반 보이스 피싱 방지 시스템을 제안한다. 통화 중 음성을 Whisper 기반 음성 인식 모델을 통해 실시간으로 텍스트로 변환하고, LLaMA 기반 언어 모델을 활용하여 대화 문맥을 분석함으로써 보이스 피싱 가능성을 판단한다. 또한 음성 신호의 오디오 특성을 분석하여 합성 음성 여부를 함께 고려하는 다단계 탐지 구조를 설계하였다. 실험 결과, 제안한 시스템은 기존 키워드 기반 탐지 방식 대비 오탐지율을 약 23% 감소시켰으며, 딥보이스 기반 보이스 피싱 탐지 정확도는 약 18% 향상되는 결과를 보였다.

Ⅰ. 서 론

현대 사회는 통신 기술과 미디어의 급속한 발전으로 시간과 장소의 제약 없이 누구와도 즉시 소통할 수 있는 환경이 조성되었다. 이러한 기술 발전은 일상생활의 편의성을 크게 향상했지만, 동시에 이를 악용한 범죄 또한 지속적으로 증가하고 있다. 그중 대표적인 범죄 형태가 전화 통화를 이용한 보이스 피싱이다. 기존의 보이스 피싱은 사람이 직접 전화를 걸어 금전이나 개인정보를 요구하는 방식이 주를 이루었으나, 최근에는 인공지능 기술을 활용한 새로운 유형의 보이스 피싱이 등장하고 있다. 특히 특정 인물의 음성을 학습하여 자연스럽게 합성하는 '딥보이스(Deep Voice)' 기술을 이용한 보이스 피싱 사례가 증가하고 있다[1].

딥보이스 기반 보이스 피싱은 가족, 지인, 공공기관 관계자의 목소리를 사실적으로 모방할 수 있어 피해자가 이를 실제 인물의 목소리로 오인할 가능성이 높으며, 기존 방식보다 더욱 교묘하고 위험한 범죄로 인식되고 있다. 기존 보이스 피싱 탐지 연구들은 주로 텍스트 기반 키워드 분석이나 규칙 기반 탐지에 의존해 왔다[1]. 그러나 이러한 방식은 문맥을 충분히 반영하지 못하고, 일상적인 대화에서도 특정 키워드가 포함될 경우 오탐지가 발생하는 문제가 있다. 또한 딥보이스와 같은 음성 합성 기술을 활용한 범죄를 효과적으로 탐지하는 데 한계가 있다.

이에 본 논문에서는 기존 보이스 피싱 탐지를 넘어, 딥보이스 기반 보이스 피싱까지 탐지할 수 있는 AI 기반 보이스 피싱 방지 시스템을 설계 및 구현하고자 한다. 음성 인식, 자연어 처리 기반 문맥 분석, 키워드 기반 분석, 오디오 특성 분석을 결합한 통합 구조를 제안하며[2], 이를 모바일 환경에서 실시간으로 활용 가능한 애플리케이션 형태로 구현하는 것을 목표로 한다.

Ⅱ. 본론

본 논문에서 제안하는 시스템은 음성 입력 → 텍스트 변환 → 문맥 및 키워드 분석 → 오디오 특성 분석 → 위험도 산출의 다단계 구조로 설계되었다. 사용자의 통화 음성은 실시간으로 수집되며, 일정 길이의 음성 데이터 단위로 분할되어 분석한다.

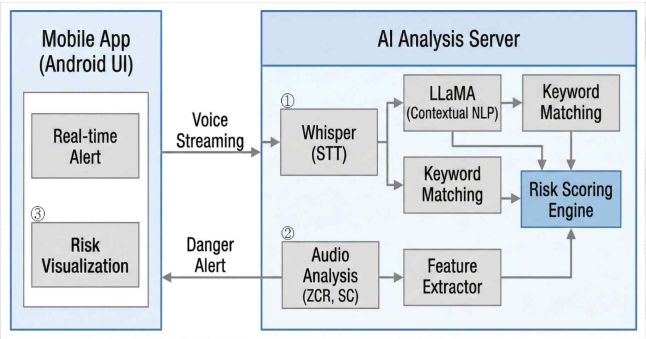


그림 1. 전체 시스템 구조.

먼저 수집된 음성 데이터는 Whisper 모델을 이용하여 텍스트로 변환된다[2]. Whisper는 잡음 환경에서도 비교적 안정적인 음성 인식 성능을 보이며, 실시간 처리에 적합하다는 장점이 있다. 변환된 텍스트는 이후 자연어 처리 단계에서 문맥 분석의 입력으로 사용한다.

문맥 분석 단계에서는 LLaMA 기반 언어 모델을 활용하여 단순 키워드 매칭이 아닌 대화의 의미와 흐름을 고려한 보이스 피싱 가능성을 평가한다. [3] 동시에 경찰청 등 공공기관에서 제공하는 보이스 피싱 관련 의심 키워드를 활용한 규칙 기반 분석을 수행하여 문맥 분석 결과를 보완한다[4]. 또한 딥보이스 탐지를 위해 오디오 특성 분석을 추가로 수행한다. 이를 통해 텍스트 분석만으로는 탐지하기 어려운 음성 합성 기반 공격에 대응할 수 있도록 하였다. 최종적으로 각 분석 결과를 종합하여 위험도를 산출하고, 일정 기준 이상일 경우 사용자에게 실시간 경고를 제공한다.

그림 1에서 ②은 딥보이스 기반 보이스 피싱 탐지를 위해 본 논문에서 오디오 신호의 특성 분석을 수행하였다. 대표적으로 Zero Crossing Rate (ZCR)와 Spectral Centroid (SC)를 활용하였다[5].

ZCR은 음성 신호가 0을 교차하는 빈도를 나타내며, 음성의 진동 특성을 반영한다. 일반적으로 합성 음성은 자연 음성과 비교했을 때 ZCR 값이 일정한 패턴을 보이는 경향이 있다. Spectral Centroid는 음성 신호의 주파수 분포 중심을 나타내며, 합성 음성인 경우 특정 주파수 대역에 에너지가 집중되는 특성을 보인다.

본 논문에서는 이러한 특성값을 분석하여 딥보이스 가능성을 판단하였으며, 분석 결과를 문맥 기반 위험도 점수에 가중치 형태로 반영하였다[2] .이를 통해 딥보이스가 사용된 경우 위험도가 추가로 상승하도록 설계하였다.

그림 1-③에 해당하는 최종 위험도는 문맥 기반 분석 점수, 키워드 기반

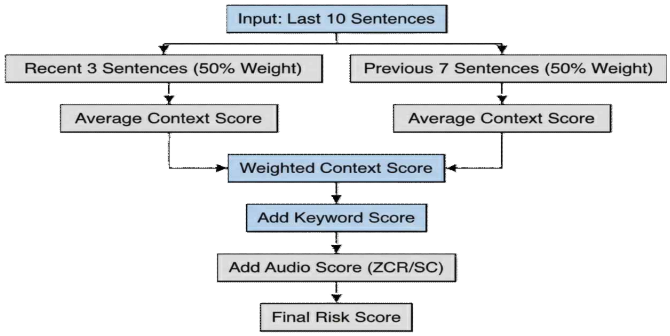
점수, 오디오 특성 분석 결과를 종합하여 산출한다.

시스템은 일상적인 대화로 판단되는 경우 키워드 기반 가중치를 최소화하여 오탐을 줄이고, 고위험 문맥으로 판단될 경우 가중치를 강화하는 방식으로 설계하였다.

위험도 산출 과정에서는 최근 대화 흐름을 반영하기 위해 총 10개의 최신 문장을 기준으로 분석을 수행한다. 이 중 최근 3개 문장의 분석 결과에 50%, 이전 7개 문장의 분석 결과에 50%의 가중치를 부여하여 최종 위험도를 계산한다. 이를 통해 단일 문장에 의한 일시적인 판단 오류를 완화하고, 지속적인 위험 패턴을 효과적으로 반영할 수 있도록 하였다.

또한 기존 키워드 기반 보이스 피싱 탐지 방식과 비교하여 키워드 가중치를 최소화하고, 오디오특성 분석을 통해 자연 음성으로 판단되는 경우 위험도를 추가로 감소시켰더니 정상 대화에 대한 오탐지율을 약 23% 감소시켜 최종 오탑률이 95% 이상의 효과를 보았으며 딥보이스 기반 보이스 피싱 탐지 정확도는 약 18%향상시켜 최종 정확도 91%의 효과를 보였다.

Algorithm 1. 위험도 산출 알고리즘.



또한, Algorithm 1. 위험도 산출 알고리즘을 이용하여 위험도가 연속적으로 상승하는 경우 위험도 증가 폭을 점진적으로 키우고, 반대로 위험도가 지속적으로 감소하는 경우 감소 폭을 키우는 방식으로 가중치를 조정하였다. 이러한 구조를 통해 보이스 피싱 특유의 점진적인 위험 증가 패턴을 효과적으로 탐지할 수 있도록 하였다.

그림 2. 애플리케이션의 기본앱 설.정

그림 2는 딥보이스 피싱 방지 애플리케이션을 기존 전화 애플리케이션 대신 기본으로 설정할 수 있는 그림이다. 통화 중 백엔드 부분에서는 음성을 Whisper 텍스트로 변환 후, LLaMa가 문맥을 파악한 후 파악한 문맥을 통해 산출된 위험도와 키워드 분석을 통해 산출한 위험도를 통합해 최종 위험도를 나타낸다.

그림 3. 일상적인 대화 결과.

그림 3의 대화는 일상적인 대화 문맥과, 의심 키워드가 없는 대화를 분석한 결과를 보여준다. 해당 대화에는 보이스 피싱과 관련된 의심 키워드가 포함되어 있지 않으며, 음성 신호 분석 단계에서 실행되는 ZCR과 SC 기반 딥보이스 판별 결과 또한 정상 음성으로 판단되었다. [5] 이에 따라 시스템은 해당 대화를 정상 통화로 분류하고, 최종 위험도를 0.0%로 산출하였다.

이는 대화 내용과 음성 특성 모두에서 보이스 피싱 징후가 발견되지 않은 경우 낮은 위험도가 유지됨을 의미한다. 반면, 대화 내에 보이스 피싱 관련 키워드가 포함되거나 ZCR 및 SC 분석을 통해 딥보이스 가능성이 탐지될 경우는 위험도가 50% 이상으로 증가하도록 설계하였다.

그림 4. 보이스 피싱 대화 결과.

그림 4는 보이스 피싱 의심 대화 문맥이 입력된 경우의 분석 결과를 나타낸다. 해당 대화에는 금융 정보 요구, 긴급 상황을 가장한 표현 등 다수의 보이스 피싱 의심 키워드가 감지되었으며, 이에 따라 텍스트 분석 단계에서 높은 위험 점수가 부여되었다. 그 결과, 시스템은 해당 통화를 보이스 피싱 가능성이 높은 사례로 판단하고 사용자에게 경고를 제공한다.

Ⅲ. 결론

본 논문에서는 음성 인식, 자연어 처리, 오디오 특성 분석을 결합한 딥보이스 기반 보이스 피싱 방지 시스템을 설계하고 구현하였다. 제안한 시스템은 기존 키워드 중심 탐지 방식의 한계를 보완하여 문맥 기반 판단과 딥보이스 탐지를 동시에 수행할 수 있도록 설계되었으며, 모바일 환경에서도 실시간으로 활용 가능한 구조를 제안하였다. 실험 결과, 문맥 분석과 오디오 특성 분석을 결합한 방식이 기존 방식 대비 오탐지율을 감소시키고, 딥보이스 기반 보이스 피싱 탐지 성능을 유의미하게 향상함을 확인하였다.

향후 연구에서는 더욱 다양한 오디오 특성 지표와 딥러닝 기반 음성 합성 탐지 모델을 적용하여 탐지 정확도를 추가로 향상하고, 오프라인 환경에서도 동작 가능한 경량화 모델을 적용할 예정이다.

참 고 문 헌

[1] Z. Wu et al., “Spoofing and Deepfake Speech Detection: A Survey,” IEEE Signal Processing Magazine, 2021.

[2] A. Radford et al., “Robust Speech Recognition via Large-Scale Weak Supervision,” ICML, 2023.

[3] H. Touvron et al., “LLaMA: Open and Efficient Foundation Language Models,” arXiv, 2023.

[4] 경찰청, “보이스 피싱 범죄 유형 및 예방 가이드,” 2023.

[5] T. Giannakopoulos, “Introduction to Audio Analysis,” Academic Press, 2014.